Universidad Autónoma de Madrid
Faculty of Sciences
Instituto Universitario Nicolás Cabrera

Development of a hybrid computational method for the study of large-scale conformational transitions in proteins

Doctoral Thesis

Ilaria Mereu

Madrid, October 24, 2014

Instituto Universitario Nicolás Cabrera
Faculty of Sciences
Universidad Autónoma de Madrid

Development of a hybrid computational method for the study of large-scale conformational transitions in proteins

Doctoral Thesis

Presented by Ilaria Mereu
under the supervision of
Dr. Andrea Nicole Dölker and Prof. Francesco Luigi Gervasio

Madrid, October 24, 2014

ACKNOWLEDGMENTS

This thesis has been possible thanks to many people, and here is just a minimal list of them. I owe them part of this step towards the aspiration to a career in theoretical biological physics, an aspiration that has no professional equivalent for me.

I am thankful for the many chances to learn that I have been offered by the directors of my group and program Francesco Gervasio and Alfonso Valencia. I also gratefully thank: Nicole Dölker, you have been a great teacher, role model, and you created key space and balance in the workplace thanks to your natural and peaceful flair. My official and unofficial PhD commission: Guillermo Montoya, Juan Fernández Recio, Michael Tress and Daniel Lietha, for helping me see how to maximize the potential of my ongoing work. Alfonso Valencia and David Pisano for the computational infrastructure. Marta Camacho, Angel Carro, Eduardo León, Giorgio Saladino, Belén Baneres, Jorge Valencia, José M. Fernández, Paz Bardaji and David Carbonell in CNIO and Raúl Guantes and Manuela Moreno at UAM, for useful technical suggestions and collaborative attitude. Simone Marsili and Antonio S. Torralba, for intellectual and technical contributions and for the example of your work ethic and aesthetic. Mohamad-Ali Fawal for always having a positive word for me. Maria Morando, Jason Morgan, Txema González-Izarzugaza and Paolo Pani for good and unselfish advice. Ludovico Sutto for technical hints. The CNIO for proposing prestigious and thought-provoking seminars.

A heartfelt acknowledgment goes to: Matteo Bianchi, whose attention and assertiveness brought a remarkable difference for contents, timing and clarity. Raffaela Cabriolu, since we share much of our background, path, struggles and dreams, and you have been like a sister for me. Daniel Rico, who will form many well-rounded minds thanks to his mentoring talent and solid foundations. Olivia Garandeau, Jenifer Clausell and Marta Camacho, three women I pride myself having grown a bit with, during these years.

Questa tesi è stata possibile grazie a molte persone, e qui se ne trova solo una lista essenziale. Devo loro parte di questo passo verso l'aspirazione a una carriera in fisica biologica teorica, un'aspirazione che non ha equivalenti professionali per me. Le prime tra queste persone sono mia madre e mio padre. Questa tesi è dedicata a loro, ed è stata possibile grazie al loro supporto e alle azioni che hanno sempre intrapreso per aiutarmi a riconoscere ed esprimere le qualità che hanno visto in me.

Y muchas gracias a España. A su estado social y a todas las entidades públicas y privadas que me han transmitido su apoyo y han confiado y creído en mí. A todas las personas de la Obra Social La Caixa y en particular, a don Juan María Nin y a doña Rosa Maria Molins, por invertir en esta forma de desarrollo y la integridad con la que mantuvieron esta elección. Espero tener la oportunidad de revertir a esta tierra y sociedad la gran confianza que me ha brindado.

CONTENTS

PRESENTACIÓN

En la presente tesis, utilizamos simulaciones de dinámica molecular (DM) clásica y potenciales basados en la estructura para estudiar cambios conformacionales de gran escala en proteínas. Se propone un esquema computacional capaz de reproducir este tipo de desplazamientos y que permite realizar cálculos de energía libre compatibles con las capacidades actuales de cómputo.
La posibilidad ofrecida hasta la fecha por los métodos de biofísica computacional es simular hebras macromoleculares como dominios de proteínas quinasa determinados con alta precisión física y en detalle atómico. Sin embargo, estos métodos presentan importantes limitaciones, que reducen los sistemas descriptibles a estructuras largas de apróximadamente 300 residuos (teniendo en cuenta recursos computacionales estándar para cálculos de energía libre). Para superar estas limitaciones, añadimos a nuestro potencial un término basado en la estructura. Este término reduce el coste computacional de la simulación, si bien supone una pérdida de precisión. Dicha innovación metodológica permite alcanzar la reproducción de transiciones conformacionales de gran escala y posibilita la extracción de datos de energía libre. Específicamente, nos permitió la simulación de una cadena de la proteína quinasa c-Abl de hasta 450 residuos.

En los siguientes capítulos describiré cómo se implementó este esquema computacional y los resultados que se obtuvieron con el mismo. En el primer capítulo presento los conocimientos previos que inspiraron el modelo. En el segundo capítulo describo como se implementó este esquema, tratando la relación entre pérdida de precisión y ganancia en eficiencia, y en el tercero su validación en el plegamiento de pequeñas proteínas. Concluyo ilustrando en el cuarto capítulo la aplicación de este protocolo en el caso de la proteina c-Abl y, finalmente, los resultados obtenidos con este sistema y la discusión de los mismos.

1

INTRODUCTION

1.1 CONFORMATIONAL TRANSITIONS IN PROTEINS

In this work, we use classical molecular dynamics simulations and structure-based potentials to study large-scale conformational changes in proteins. We propose a computational scheme able to reproduce such motions within the current computational capabilities that can be used for free energy calculations.

The interest of this investigation is both clinical and fundamental. On the one hand it contributes to the mechanistic understanding of the regulation of multidomain proteins whose dysregulation is associated to human cancer like in the case under study, the c-Abl protein kinase. On a more foundational front, this protocol provided us with the chance to investigate protein conformational changes ranging from angstroms to tenths of nanometers, rarely accessible to current average computational capabilities, reaching a closer depiction of protein dynamics and mechanicism at the mesoscopic scale.

Proteins are a fundamental category of macromolecules within the realm of living matter. They are responsible of performing a plethora of functions that are indispensable to the life of cells and organisms. Among them, enzymes are catalyst proteins: they lower the activation energy of a specific biochemical reaction, increasing its rate without being consumed by it. For many enzymes, scientists were able to identify amino acids directly involved in catalytic activity (for an example, see [1]). This led to the notion of associating the phases of an enzyme's activity and regulation to specific structural conformations.
Computational methods of atomistic resolution such as molecular dynamics (MD, [2, 3]) can be extremely valuable in characterizing these conformations and help to quantify the accessibility of the conformational states involved in enzymatic function.

In a typical framework of classical MD, a protein is represented as a set of

N bound atoms whose position is specified by $3N$ coordinates and evolves in time according to a physically accurate potential. Different conformations occupy different positions in the $3N$ dimensional free energy space and are separated by barriers, whose disparate heights define a landscape. Molecular simulations (MD and derived enhanced sampling methods - for reference see, for example, [4]) are capable of estimating the free energy profile along generalized coordinates chosen from this multidimensional landscape and to provide insights on the structural elements involved in conformational barrier crossing and, more generally, into protein function.

Because of the value they add to the description of living matter at the atomistic level, molecular simulations are now recognized as powerful tools for the understanding of any phenomenon where biomolecules are involved. The sharpness of their vision greatly helps in cases that see subtle deviations from the wild-type protein at play, like clinical mutants. In fact, the polypeptide chain that builds up a wild-type protein can be modified by single residue mutations or larger erroneously transcribed strands, or both, and in some cases develop in the molecular mechanism governing a malignant cell. As will be described briefly in the next section, this is the case of c-Abl protein and of its variations that drive certain forms of leukemia: the Bcr-Abl fusion protein and its clinical mutants.

The main task of molecular simulations is to generate the ensemble of conformations that corresponds to a specific system and set of thermodynamic conditions in agreement with the laws of statistical mechanics. Once a reliable ensemble has been generated by the simulation, new knowledge about the system can be gained through the ensemble's extremum principle: a thermodynamic law involving an observable such as entropy and free energy whose maxima or minima (extrema) indicate the states of equilibrium and how the system most likely tends to reach them. For transitions in the canonical ensemble (NVT), the extremum principle refers to the Helmholtz free energy, so this quantity becomes cardinal to the characterization of a system at equilibrium. Moreover, being the free energy a peculiar property of the specific system, any modification in the protein and its environment must be reflected on its free energy landscape. Consequently, perturbations in the physiological free energy profile can be exploited as a detection tool, as they may indicate important alterations of the equilibrium state and be part of the molecular picture behind a pathologically misbehaving protein.

1.2 THE CASE OF C-ABL

The c-Abl protein is a non-receptor tyrosine kinase, a modular signaling protein involved in cell differentiation, cell division, cell adhesion and stress response [5]. As a kinase, it catalyzes phosphate transfer onto protein substrates. Its catalytic domain is tightly regulated through the interplay of several structural intra- and interdomain interactions that determine its degree of autoinhibition, and exhibits low constitutive catalytic activity due to tight regulation [6].
Bcr-Abl is associated to more than 90% of the cases of chronic myelocytic leukemia, CML [7], and to one third of the cases of acute lymphocytic leukemia, ALL [8]. It is a chimeric protein encoded by an abnormal chromosome (the so-called "Philadelphia chromosome") that includes only part of the gene that expresses c-Abl. The consequence is that Bcr-Abl shares with c-Abl the catalytic domain, but not the rest of its endogenous regulatory domains [9]. It exhibits constitutive catalytic activity and is considered necessary for inducing cell transformation and cause malignancy [10].
From a free energy landscape standpoint, Bcr-Abl can be interpreted as a system that, in its irregular layout, admits differently distributed populations of active and inactive conformations of the catalytic domain with respect to its physiological counterpart c-Abl and, because of this, reproduces the signaling effect pertaining to c-Abl in aberrant proportions. Full understanding of the molecular reasons behind this ability of Bcr-Abl is a subject of research and would greatly help the design of inhibitors for the treatment of different forms of leukemia in humans. Gaining quantitative information on which portions of the protein and of the free energy landscape in c-Abl (and indirectly of Bcr-Abl) are correlated and relevant for function is currently a highly non-trivial endeavor and is one of the subjects this thesis addresses.

The c-Abl kinase protein is not an isolated case for its role in cancer development among the kinase family.
All protein kinases act as switches by catalyzing phosphoryl transfer from ATP to protein substrates [11]. Because of their power to modify substrate activity through phosphorylation (the most abundant form of post-translational modification and cellular regulation in eukaryotes), they intervene in most of the signal transduction processes in eukaryotic cells. Additionally, they participate

in metabolism, transcription, cell cycle progression, and other cellular vital functions. They turn on and off cell signaling pathways and for this reason the balanced regulation of their catalytic domain - realized by differentiated sets of adaptor proteins - is of utmost relevance [12]. Their action is counterbalanced by the family of phosphatases, in charge of removing the phosphoryl group from phosphorylated residues.
Soon after the publication of the sequence of the human genome [13], and thanks to the availability of several information sources for human sequences, a number of 518 putative protein kinase genes was estimated via sequence analysis, accounting for about 1.7% of the entire human genome [11]. For their strategic role in cytoplasmic signaling, kinases can often represent vulnerable points of their signaling pathway, and their drift from a properly regulated condition can assume pathological consequences.

For their role of switches, kinases not only are vulnerable to pathological dysregulation but, at the same time, offer the chance of a possible therapy. In fact, targeting a malignant kinase for selective inhibition has demonstrated highly promising over the last fifteen years.
A brilliant example of this strategy in the history of pharmacological research is exactly Bcr-Abl with its first inhibitor, the cancer drug imatinib (also known as Gleevec or Glivec). The approach behind imatinib pharmacological development was one of rational drug design [14], targeted to specifically inhibit the Bcr-Abl fusion protein at the active site. Imatinib has shown IC_{50} values ranging between 0.1 and 0.5 μM for c-Abl, v-Abl and Bcr-Abl kinase inhibition and three orders of magnitude lower values on other closely related kinases such as c-Src, EGFR, c-Lyn, PKA and others [15]. Soon after running clinical trials that resulted in minimal adverse effects and complete hematologic remission in 53 over 54 patients [16], the pharmaceutical company Novartis announced the United States FDA approval for imatinib through oral therapy for the treatment of patients with chronic myeloid leukemia (CML). Along with its second-generation inhibitors [17, 18], the rationally designed inhibitor imatinib is now a clinical reality. In addition, its efficacy represents a powerful proof of concept for rational drug design as an effective pharmacological strategy, when tailored to molecular abnormalities associated to a human disease. Yet, mutations in the kinase domain often give rise to drug resistance at later stages of the disease, so these inhibitors do not represent the definitive answer

yet. They are necessary but not always sufficient to keep the disease under control. Thus, the need for fully unraveling the molecular regulation mechanism of this molecule remains, and with it the hope of finding a conclusive set of inhibition strategies, possibly including allosteric sites. Current inhibitors, in fact, all bind at the active site. Thus, most of them are rendered inactive by the same resistance mutations. Therefore, inhibitors targeting allosteric sites would be of great value [19].
Considered the large physiological impact of the kinase family, the potential of selective kinase inhibition is still far from depletion and holds the promise of new breakthroughs [20].

The chance offered today by computational biophysics methods is to simulate macromolecular strands like specific kinase domains with high physical accuracy (through classical force field MD) and in atomistic detail. A computational insight of this kind can help ground experimental choices in physicochemical bases and save time and expensive and complex experimental efforts for the most promising routes. In the following chapter, I will describe the computational strategies we applied with the aim of providing an efficient and comprehensive structural intepretation for the malignant drift of c-Abl.

1.3 COMPUTATIONAL LIMITS

Although state-of-the-art computational simulations for biomolecules have given important insights, there is currently a series of strong limitations to the results within reach of these methods.
One limiting factor is the energy barriers' height that hinder conformational sampling. In classical or replica exchange MD (REMD from now on) simulations, that are based on Boltzmann sampling, the only chance a system has to overcome a barrier is an energy fluctuation, whose probability decreases exponentially with the barrier's height. Thermally high barriers often shield phenomena of biological interest but in the community their crossing is defined, for its technical difficulty, a "rare event". Practically, just this notion severely limits the phase space accessible to a simulated system within the barriers whose height it is unlikely to repeatedly overcome.
Beyond the ruggedness and the extent of the landscape to sample, a second notion to consider when designing a computational experiment lies in the combination of the total computational power at disposal and the system size. The

computational power, namely the simulation time per CPU available to the user for her simulations, is the most tangible practical limit to every in silico experiment. Typically, in a standard computer cluster like the ones at disposal for most of the present work, computational power allows - at optimal scaling - to simulate about a thousand atoms per CPU for one microsecond in a few months [21]. Notable exceptions to this average estimate exist. They produced [22, 23, 24] and will keep producing new remarkable insights on longer timescales, but they remain uncommon.

Since convergence of the sampling to equilibrium is required for drawing any grounded conclusion and computational load is the principal practical limit, care has to be given for convergence to be reached in a reasonable time for a given system size. These limitations concur to set quite restrictive limits to the phenomena accessible to molecular simulations with classical force fields.

For these reasons, to achieve the aim of simulating large-scale conformational transitions on proteins and protein complexes much larger than 300 amino acids (the approximate size of the kinase domain) with general-purpose computational resources, classical force field MD or REMD simulations with explicit solvent are not currently a viable choice. These methods rely, for the future, on the advance of computational power to afford such goal.

In the meantime, the request for methodological alternatives encouraged the exploration of new ways to tackle the rare events problem. Two possible paths leading to easier access to such events are: 1) acting on the front of the sampling method, going beyond Boltzmann sampling, and 2) acting on the front of the potential energy function describing the interactions between atoms, by reducing the force field complexity.

In this work, metadynamics [25] was the enhanced sampling method of choice. Its implementation PLUMED [26], provides the user with the tools to define a reduced number of generalized coordinates in the free energy landscape and drive the system over the transition barriers along these coordinates. These tools and the combination with a replica exchange scheme have been a fundamental ingredient for succeeding in the simulation of c-Abl's large-scale transitions.

Still, although enhanced sampling methods increase the accessible configuration space, the size of the system and the amplitude of the conformational transitions are still a limiting factor. To overcome it, we combined the most powerful sampling methods at our disposal with a boost at the force field level,

by combining a structure-based term with a state-of-the-art classical force field. The next chapter will describe this protocol in detail, preceded by some background notions that wish to help situate the different approaches involved.

The model was first tested and tuned to the folding transitions of two small proteins, one of which rich in α helices and the second in β sheets: the villin headpiece HP35 and c-Src SH3 domain. This was made in order to calibrate the free parameter of the model (the energy per native contact, ε, in kJ/mol) on a representative benchmark, so to ensure physical reliability in relation to the description of the two major structural elements occurring in protein structure. Chapter 3 describes and discusses this validation phase of the work.

Chapter 4 illustrates the application of the protocol to c-Abl, by first introducing in detail the molecular features most relevant to this work, and finally the results obtained and their discussion.

2

COMPUTATIONAL METHODS

Large-scale conformational transitions in proteins still remain beyond the reach of unbiased explicit-solvent MD simulations. Since this was our ultimate target, many measures had to be orchestrated to hit it. Not only the structure-based nuance was a fundamental character of our pipeline, but also enhanced sampling methods were required. At this level, only parallel tempering metadynamics (PTmetaD) was capable of attaining meaningful results on the application front.
This chapter will describe the methods employed in this thesis to study folding and large-scale conformational transitions. It starts from a brief description of molecular dynamics and the other enhanced sampling methods we employed to overcome current sampling limitations. The next part illustrates the background and the reasons of structure-based modeling. Finally, it reports the methodological details of the structure-based model that this thesis presents.

2.1 SAMPLING METHODS

2.1.1 MOLECULAR DYNAMICS

In this work, proteins and peptides (and, in some cases, their solvent too) are treated according to classical molecular dynamics, namely as fixed sets of N bound atoms whose position is specified by $3N$ coordinates. These coordinates collectively define a conformation, and it is assumed that no bond is created or broken during the simulation. The technique of all-atom classical molecular dynamics estimates numerically the forces acting on atoms and the consequent displacements they induce by assuming a potential energy function partitioned in five terms, three referred to interactions between bonded atoms and two rel-

ative to nonbonded interactions, computationally more expensive.

$$
\begin{aligned}
U &= \underbrace{U_{\text{bonds}} + U_{\text{angles}} + U_{\text{dihedrals}}}_{\text{bonded interactions}} + \underbrace{U_{\text{LJ}} + U_{\text{electrostatics}}}_{\text{nonbonded interactions}} \\
&= \sum_{i,j \in b} \frac{1}{2} K_{b,ij} (r - r_0)^2 + \sum_{i,j \in b} \frac{1}{2} K_{\vartheta,ij} (\vartheta - \vartheta_0)^2 \\
&\qquad + \sum_{i,j \in b} \sum_{n} K_n [1 - \cos(C_n \varphi + \delta_n)] \\
&+ \sum_{i,j \in nb} \varepsilon_{ij} \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right] + \underbrace{\sum_{i,j \in nb} K_C \frac{q_i q_j}{r_{ij}}}_{\text{partial charges}}
\end{aligned}
\tag{2.1}
$$

where, for the atom pair i,j:

- $K_{b,ij}$ is the force constant and r_0 the equilibrium distance for the bond of length r;
- $K_{\vartheta,ij}$ is the force constant and ϑ_0 the equilibrium angle for the angle ϑ;
- the dihedral angle φ is modeled as a periodic function with n barriers defined by the parameters K_n, C_n and δ_n;
- ε_{ij} is the depth of the well placed at r_{ij}^0 for the Lennard-Jones interaction at distance r_{ij};
- q_i and q_j are the respective partial charges at distance r_{ij} and K_C is the electric conversion factor over the dielectric constant.

While the computational biophysics community mostly uses this functional form, the parameters vary from one force field to another. The force field's coefficients depend on the nature of atoms and bonds involved and are derived by fitting on ab-initio calculations and experiments. In the effort of establishing a reference for the community, different state-of-the-art force fields (Amber ff03 [27], Amber ff99SB*ILDN [28, 29, 30], CHARMM22 [31] with the CMAP backbone correction [32], and a version of CHARMM22 with modified backbone torsion potentials) were recently tested on a common benchmark, the folding

of a small helical protein, with the exceptional computing means available at D.E. Shaw Research [33]. The comparison led to conclude that the overall performances of these recently developed force fields under test were consistent and successful in recovering the correct native state and folding rate. Yet, space for improvement remained in many respects.
Given a fundamental discretized time unit called *time step*, forces in MD are calculated from the classical Hamiltonian via Newton's second law of dynamics, the positions are updated and forces recalculated, recursively. Thus, MD provides dynamical information for the system under study in compliance with the fixed ensemble variables. These conditions, in turn, are enforced through constraints, thermostats or barostats that affect the computing of forces. Furthermore, if a trajectory conformed to the ergodic hypothesis (that all accessible microstates were equiprobable during the sampling), the average over the trajectory could be assumed equivalent to the average on the phase space ensemble, which gives to MD the feature of yielding also ensemble averages, besides the dynamical information.
The current standards for systems sizes and investigated biological phenomena are the result of the dominant practice in the community to use MD engines from software simulation packages for biomolecular simulation on parallel machine. Popular packages are AMBER [34], CHARMM [35], NAMD [36, 37] and GROMACS [38, 39, 40, 41].

2.1.2 PARALLEL TEMPERING

Parallel tempering is a sampling method in which a set of N temperatures is assigned to different molecular dynamics simulations of the same system. It is a form of replica exchange MD (REMD). The rationale for accelerating standard MD sampling here is that high-temperature replicas are capable of crossing energy barriers more easily than the low temperature ones, which in turn are more likely to sample the local minima they happen to occupy. In parallel tempering, a number N of MD simulations are started in parallel at temperature $T_1 < T_2 < \cdots < T_N$. Periodically, an exchange of coordinates is attempted between a neighboring pair of configurations i, i+1. The acceptance probability of this move is:

$$P(1 \leftrightarrow 2) = \min\left(1, \exp\left[\left(\frac{1}{k_B\, T_1} - \frac{1}{k_B\, T_2}\right)(U_1 - U_2)\right]\right) \tag{2.2}$$

This yields an overall enhancement of efficiency in the sampling, which remains of Boltzmann type [42, 43].

2.1.3 METADYNAMICS

One of many methods aiming to provide a solution to the rare events problem, metadynamics (metaD from now on, [25]) detaches from Boltzmann's sampling by applying to a system's potential an additional time-dependent term on a subset of the total degrees of freedom.
Metadynamics' core idea [44] finds its roots in the concepts of dimensionality reduction (pioneered by Y. Kevrekidis and coworkers [45, 46]) and shares some similarities to tabu search, umbrella sampling, conformational flooding and the local elevation method [47]. In its scheme, a subset of the whole free energy space is defined by the choice of a set $\{\mathbf{s}\}$ of a reduced number of generalized coordinates (usually referred to as collective variables, CVs).
To understand the basic picture, let us first consider a definition that turns fundamental in the realm of free energy calculations. In an unbiased MD simulation for a canonical system, the Helmholtz free energy is defined, up to an additive constant, by

$$F(\mathbf{s}) = -\frac{1}{\beta} \ln N(\mathbf{s}) \tag{2.3}$$

where $N(\mathbf{s})$ is the number of visited configurations at the point $\mathbf{s}$ of the CV space. These CVs are functions of structural elements of the system, and represent the transition's order parameter (so to concisely describe and discriminate between different configurations for the event under investigation). In the actual cases examined in the following of this thesis, the CVs chosen were distances (between either single atoms or centers of mass) and contact map-based s-paths (see below for more details).
The metaD potential contains a bias term that is built adaptively at regular time intervals. In the spirit of umbrella sampling [48], metaD generates the ensemble associated to a known bias $V(\mathbf{s})$ and recovers the free energy through the formula

$$F(\mathbf{s}) = -\frac{1}{\beta} \ln N(\mathbf{s}) - V_{bias}(\mathbf{s}) \tag{2.4}$$

At a fixed time stride τ_G, an N-dimensional gaussian is added to the potential at the system's coordinate, in order to force it to move away from the local minimum of the N-dimensional free energy landscapes where it is placed in. At time t, the metaD bias potential has the following form:

$$V_{bias}(\mathbf{s}, t) = \int_0^t dt' \omega \exp\left(- \sum_{i=1}^{d} \frac{(s_i(q) - s_i(q(t'))^2}{2\sigma_i^2} \right), \tag{2.5}$$

where d is the number of collective variables, σ_i is the width of the gaussian in the i-th direction and ω has the dimension of an energy rate. The parameters are strongly system dependent and must be assigned opportunely by the user.

This mechanism enhances the probability of barrier crossing from exponential (as in Boltzmann's sampling) to linear in the barrier high. The response of the system along the trajectory to the bias potential builds up the time-dependent, landscape-adaptive potential, provided the systems has the time to diffuse towards the new free energy minimum (gaussians have to be deposited at moderate stride to ensure this). At convergence the system will have eventually escaped any minimum it may be trapped in, and in the case of well-tempered metadynamics it can be mathematically proved that the underlying free energy profile is reconstructed by integrating the bias potential [49].

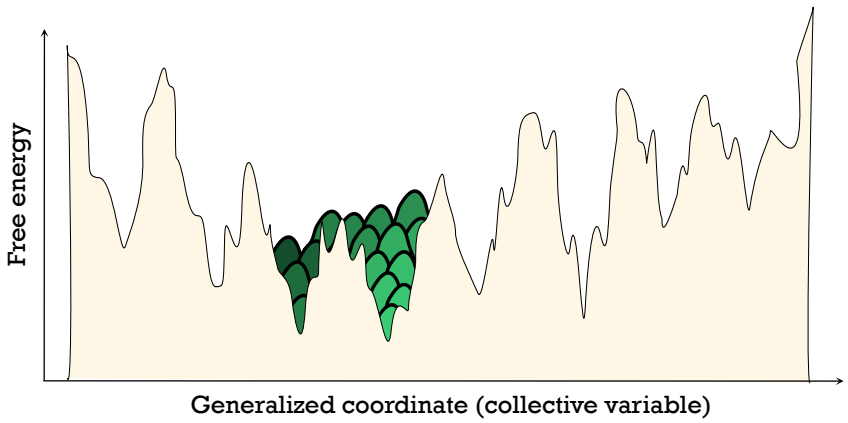

Figure 1: Mechanism of gaussian hills addition in metaD.

The computational implementation of metadynamics was provided by the publicly available PLUMED software package [26]. PLUMED also implements several auxiliary developments that were designed to enhance the basic and powerful sampling scheme described above. Here we mention and synthetically describe in principle those involved in the original work we present, and redirect to the suggested references for further details.

- One of the most relevant metaD improvements we resorted to is well-tempered metadynamics [50]. The well-tempered rationale significantly refines basic metaD by defining the bias potential $V(\mathbf{s}, t)$ as a monotonically increasing function of the histogram $N(\mathbf{s}, t)$ of the CV values at the same time, thus embedding in the bias potential a term that discourages already visited configurations by construction. Specifically, the time-dependent potential of well-tempered metaD has the form:

$$V(\mathbf{s}, t) = \Delta T \ln \left(1 + \frac{\omega N(\mathbf{s}, t)}{\Delta T} \right) \tag{2.6}$$

  where ω is an energy rate and ΔT is a characteristic energy. This choice implies a damping, time-dependent factor for the height w of the deposited gaussian: $w = \omega e^{-V(\mathbf{s},t)/\Delta T} \tau_G$ and convergence of the bias potential

$$V(\mathbf{s}, t \to \infty) = - \frac{\Delta T}{\Delta T + T} F(\mathbf{s}) \tag{2.7}$$

  to the exact free energy in the long time limit. This provides the simulation with a sort of adaptive brake that prevents an idle bias to grow (and the simulation to keep running) idly over a flattened free energy surface, at variance with the original metadynamics recipe, that features a fixed height for the gaussian hills and leads to bias potentials that grow indefinitely. The PLUMED user will additionally specify the bias factor $(T + \Delta T)/T$. This result found several empirical confirmations and was analytically proven for systems driven by Langevin dynamics [51].

- Just as the REMD approach was developed for MD calculations, the same scheme was devised for metaD. This is called Parallel Tempering metadynamics (PTmetaD, [52]). Taking into account the fact that each parallel simulation i is in the act of building its own bias $V_i(\mathbf{s})$, the acceptance

probability between two adjacent replicas i and j for PTmetaD can be shown to be:

$$P(i \leftrightarrow j) = \min \left\{ 1, \exp \left[\left(\frac{1}{k_B T_j} - \frac{1}{k_B T_i} \right) (U(\mathbf{s}_j) - U(\mathbf{s}_i)) + \frac{1}{k_B T_i} (V_j(\mathbf{s}_j) - V_j(\mathbf{s}_i)) + \frac{1}{k_B T_j} (V_i(\mathbf{s}_i) - V_i(\mathbf{s}_j)) \right] \right\} \quad (2.8)$$

The result is a boost at two levels in the sampling enhancement with respect to classical MD. When used in this thesis, PTmetaD had resulted the only viable solution after reiterated tentative applications of less powerful approaches such as single metaD or REMD.

- Furthermore, when it comes to conformational transitions in proteins, a non-reductive way of describing a macromolecular transition is highly desirable. Such need was addressed with the implementation of path collective variables, adapted to metaD by Branduardi et al. [53]. PLUMED implements two kinds of path CVs, the *s-path* and *z-path*. Their advantage is to compactly represent global configurations and connect them through an ordered path based on their distance $d(\mathbf{s}_i, \mathbf{s}_j)$, where $d : X \times X \to \mathbf{R}_0^+$ is a metric on X. PLUMED offers various possibilities for the metrics d, and in this work the choice fell on the contact-map based metric. It consists in the distance between the contact matrices D for a given subset of the atoms in the system, so that $d(X_j, X_i) = ||D^{(j)} - D^{(i)}||$ with

$$D_{ab}(X) = \theta(c_{ab} - r_{ab}) w_{ab} \frac{\left(1 - (r_{ab}/r_{ab}^{(0)})^{n_{ab}}\right)}{\left(1 - (r_{ab}/r_{ab}^{(0)})^{m_{ab}}\right)} \quad (2.9)$$

where $\theta(r)$ is a step function which vanishes if $r < 0$ and $r_{ab}^{(0)}$, n_{ab} and m_{ab} are flexible parameters. The definition of a path-based CV requires also the choice of an ordered set of a number R of reference conformations, representative of the extremes and relevant intermediates of the transition to describe. Two are the variables that can be monitored along the simulation:

– The position of the system along the path:

$$s = Z^{-1} \sum_{i=1}^{N} i e^{-\lambda\, d(X_i, X(t))}$$

where

$$Z = \sum_{i=1}^{R} e^{-\lambda d(X_i, X(t))}$$

is a normalization factor and $d(t)$ defines the distance between the configuration at time t and the reference frame i in the X metric. This form ensures $s(X_i) \simeq i$.

– The position off the path:

$$z = -\lambda^{-1} \log Z.$$

The path CVs employed in this thesis were contact map-based *s-paths*.

- One last tool was fundamental to this work: the reweighting technique and its PLUMED implementation [54]. Since the well-tempered metaD bias recovers a distribution for a variable $\mathbf{R}$ that is distorted with respect to the canonical distribution, a reweighting procedure is necessary to reconstruct it. Given the original bias $V(\mathbf{s}, t)$ and the time-dependent distribution of the new variable $P(\mathbf{R}(\mathbf{s}), t)$, it can be shown that the following reweighting relation holds at every t between the desired Boltzmann distribution associated to the unbiased potential $P_0(\mathbf{R}) = e^{-\beta(U(\mathbf{R}))}/Z$ and $P(\mathbf{R}(\mathbf{s}), t)$ as yielded by metaD:

$$P_0(\mathbf{R}) = P(\mathbf{R}, t) e^{\beta(V(\mathbf{s},t)+c(t))}$$

where $c(t) = -\dfrac{1}{\beta} \log \left(\dfrac{\int d\mathbf{s}\, e^{-\beta F(\mathbf{s})}}{\int d\mathbf{s}\, e^{-\beta (F(\mathbf{s})+V(\mathbf{s},t))}} \right)$ is the time-dependent bias offset and $F(\mathbf{s})$ results from eq. (2.7). This applies both to the well-tempered metaD CVs and a generic set of variables $\mathbf{R}$ extracted ex post from the same metaD simulation (in which case the relation $\mathbf{R}(\mathbf{s})$ is known at every time t).

2.2 STRUCTURE-BASED MODELING

The protein folding problem, namely the fact that an elongated and disordered polypeptide sequence univocally folds in one structure on the ms timescale only for a small subset of all the possible polypeptide sequences, found a convenient description in the energy landscape theory [55]. According to it, only those sequences that admit a spatial configuration that maximizes the number of competing sets of favorable interactions with respect to other conformations can fold fast and efficiently. This can be rephrased as the notion that only mildly frustrated sequences can efficiently fold, where the term frustration stands for the impossibility of satisfying many different sets of favorable contacts at the same time. In this view, each possible sequence is characterized by a certain degree of frustration, but protein sequences share the quality of a funneled and relatively smoothed energy profile driving their folding process. From an evolutionary standpoint, the folding behavior of proteins can be thought as resulting from an amino acid sequence optimization under the constraints of evolutionary robustness and biological efficiency [56].
From a modeling point of view, this means that there is a physical basis in shaping the energy function (the Hamiltonian) as a funnel with the expected folded structure, known a priori, at the energy minimum. This can be achieved by the definition of effective interactions between native contacts, specific pairs of atoms that are close (within a certain criterion) in the folded, or native, structure. Assigning an energetic preferentiality to native contacts overcomes the frustration problem, and is at the base of the development of models that ignore non-native interactions and reward only the native ones. These are called structure-based potentials or Gō models. Originally they were first proposed by Nobuhiro Gō in 1975 [57, 58] and have been proved successful in many cases [59, 60]. In its most common variant, the Gō model is a coarse-grained representation of the residues, each represented by a bead centered at the location of the C_α atom, and where stabilization is achieved through native interactions only [61].
Why these potentials are efficient in reproducing protein folding from a generic unfolded state is understandable, for the native conformation is just the global energy minimum of the minimally frustrated, purely structure-based Hamiltonian. However, there is a high price to pay in accuracy with respect to classical, all-atom empirical force fields that are parametrized on quantum chemistry

calculations and experiments and give access to accurate sampling. So while structure-based models proved to be powerful tools to sample the large conformational changes that occur on long timescales, they need to be supplemented with more refined calculations to investigate the atomistic detail, since the biochemistry of the problems these methods aim to address dramatically depends on local modifications of reduced subsets of residues. This quest for a balance between efficiency and accuracy recently laid the foundation for the development, after the first Gō-model, of new purely structure-based refinements such as introducing residue heterogeneity in the C_α interactions [62] and the all-atom treatment instead of the coarse grain [63, 64, 65], all always featuring native interactions to assure stability. Ideally, reproducing a long timescale process while including the most possible complete detail on relevant atomic interactions and at the same time being less computationally expensive than empirical force fields, would be of great importance in the field of molecular simulation of biomolecular systems. Still, in all kinds of purely structure-based models there is no consideration of the realistic dynamics of atoms, as it can be yielded for systems of limited size by state-of-the-art classical force fields. This compromises the predictive power of these two opposite methods and seriously limits the remarkable potential insight they could release if properly merged together. Recently, this led to the proposal of hybrid models that [66, 67] blend a realistic full atom potential with a Gō-like potential biased towards the native state structure. The model presented in this thesis joins this last category of methods and its details will be described in the next section.

2.3 MODEL

INTERACTION POTENTIAL The hybrid structure-based model [68] used in this work hinges on a Hamiltonian for heavy atoms built merging the topology yielded by a state-of-the-art force field for the bonded terms and a contact-

based potential for non-bonded interactions. It has the following functional form:

$$
\begin{aligned}
U_{\text{hybrid}} &= \underbrace{U_{\text{bonds}} + U_{\text{angles}} + U_{\text{dihedrals}}}_{\text{bonded interactions}} + U_{\text{Gō}} \\
&= \underbrace{\sum_{i,j\,\in\, b} \frac{1}{2} K_{b,ij}(r - r_0)^2 + \sum_{i,j\,\in\, b} \frac{1}{2} K_{\vartheta,ij}(\vartheta - \vartheta_0)^2 + \sum_{i,j\,\in\, b} K_{\varphi,ij}[1 - \cos(n\varphi + \delta)]}_{\text{From classical force field}} \\
&+ \underbrace{\sum_{\substack{i<j \\ \in \\ \text{native pairs}}} \varepsilon_{\text{SB}} \left[\left(\frac{r^0_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r^0_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{\substack{i<j \\ \text{non native} \\ \text{pairs}}} \left(\frac{W_{HS}}{r_{ij}} \right)^{12}}_{\text{Gō / Structure-based}}
\end{aligned}
\tag{2.10}
$$

where the structure-based part is constituted by a sum of Lennard-Jones potentials between the native pairs, each with a minimum at energy $-\varepsilon_{\text{SB}}$ located at the *native distance* r^0_{ij}, and and a hard-sphere repulsive term for all non native pairs (for which the parameter $W_{HS} = 0.2\,\text{nm}$ is defined to avoid unrealistic overlaps between atoms). This is conceptually similar to other publicly available structure-based protocols [63, 69], but differs in not leading to perfectly funneled potential energies. Rather, it includes an element of ruggedness to the energy landscape by retaining the bonded interactions, and implementing the native bias through the replacement (except for 1-4 dihedrals, that are mantained from the original force field) of non-bonded interactions with the Lennard-Jones potentials. Electrostatics was removed from the computation, assuming its effect was included implicitely in the effective interactions. Consequently, all charges were set to zero.

DYNAMICS Systems were evolved under stochastic dynamics in vacuum with the GROMACS engine [38, 39, 40, 41]. This consists in two additional friction and noise terms in Newton's equations of motion:

$$
m_i \frac{d^2\mathbf{r}_i}{dt^2} = -m_i \xi_i \frac{d^2\mathbf{r}_i}{dt^2} + \mathbf{F}_i(\mathbf{r}) + \mathring{r}_i
$$

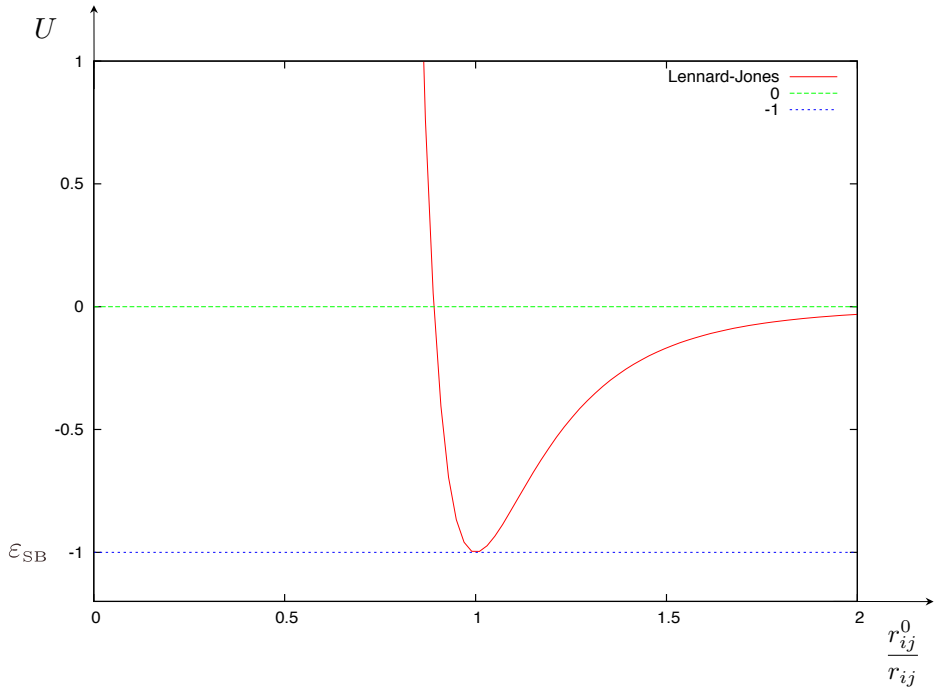

Figure 2: Lennard-Jones function, with $\varepsilon_{\mathrm{SB}} = -1$

where ξ is the friction constant $[1/ps]$ and $\mathring{r}(t)$ is a noise process with $\langle \mathring{r}_i(t)\mathring{r}_j(t+s)\rangle = 2m_i\xi_i k_B T\delta(s)\delta_{ij}$.

CHOICE OF THE CONTACT PAIRS AND MULTISTATE CHARACTER OF THE MODEL The pairs of atoms in contact (equivalently referred to in the following also as *native pairs*, *native contacts*, or just *contacts*) for the structure-based potential were chosen according to the shadow approach [64], a measure aimed to include screening effects, where two requirements are enforced: that the atoms would be included within a cutoff distance of 5 Å and a cutoff angle of 35° (see fig. 3). This is more realistic than the traditional native pair definition of C_α Gō models, which is based on a cutoff value only, as it does not allow the number of contacts to grow indefinitely, but rather leads to saturation as the cutoff distance increases. Also, pairs of atoms belonging to aminoacids separated by less than three residues in the sequence were neglected. In the case of the conformational transitions, various conformations (or native states) were

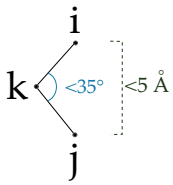

Figure 3: Rationale for determining a non-screened atom pair (i,j) in the vicinity of a screening atom k .

included in the topology, each through its set of native contacts. If a certain contact was common to various native states, the least of the two distances was chosen as the Lennard-Jones minimum in the model's potential.
The list of contacts was first extracted after equilibration with unbiased all-atom classical MD in explicit water for at least 5 ns at 310 K and further minimization in vacuum Any effect due to the presence of water molecules, either incidental or functional, has then been filtered out at the price of losing the functional ones. In the version of the model that was used for the results of the fourth chapter, such list was filtered retaining only those pairs whose average distance plus standard deviation $d + \langle d \rangle$ along 2 ns of equilibration stayed below the 5 Å threshold.

CHOICE OF ε_{SB} In the structure-based term, all Lennard-Jones functions have depth $-\varepsilon_{SB}$ at the native distance r_{ij}^0. Clearly, the behavior of the system is strongly dependent on ε_{SB}, so its choice is critical.
If ε_{SB} is too weak, the energy landscape will be excessively rugged, the system

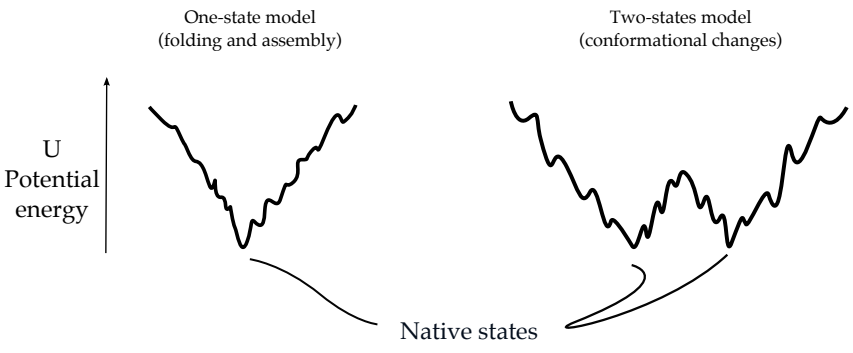

Figure 4: Potential energy profiles for a multistate model

will be mainly driven by bonded interactions and will just sample the local minimum it is trapped in. Moreover, this sampling would not even be physically accurate, because of the absence of the long-range part of the classical force field. The system would lack folding ability as well, because the essential, structure-based interactions will not be strong enough to take over the classical part of the force field. On this side of the spectrum, then, not only the model would approach too closely a classical, force field-driven molecular simulation but, for a comparable computational investment, a full classical topology would demonstrate more convenient and reliable.
If, on the other hand, if ε_{SB} is too high, folding will be achieved quickly in a mainly flattened energy landscape, but the atomistic detail of the evolution of the system will get overshadowed by the intensity of native interactions and the transition state ensemble would not be sampled long enough.

For our purposes, a range of $\varepsilon_{SB} \in (3.8, 4.5)$ kJ/mol for the coupling parameter was identified as appropriated to keep into the proper account these two opposite tendencies. To determine this range, a classical MD of 10 ns of the c-Src kinase domain in explicit solvent at 310 K (using as force field the AMBER99SB* modification of ff99SB ([29]) combined with AMBER99SB-ILDN force field [30]) was taken as reference for a series of short stochastic dynamics of 2 ns each, in vacuum at 310 K, governed by the structure-based Hamiltonian having a range of values of ε_{SB} of 2-8 kJ/mol at steps of 0.1 kJ/mol, for a total of 61 short simulations. Comparison of the root mean square deviation of the atoms' root mean square fluctuations with respect to the reference for all the heavy atoms, C_α, and sidechains' atoms was performed to assess the model's limits of optimal replication of the natural protein's flexibility. The relative plot shows a shallow minimum enclosed by the two extreme conditions in the $(3.8, 4.5)$ kJ/mol region.

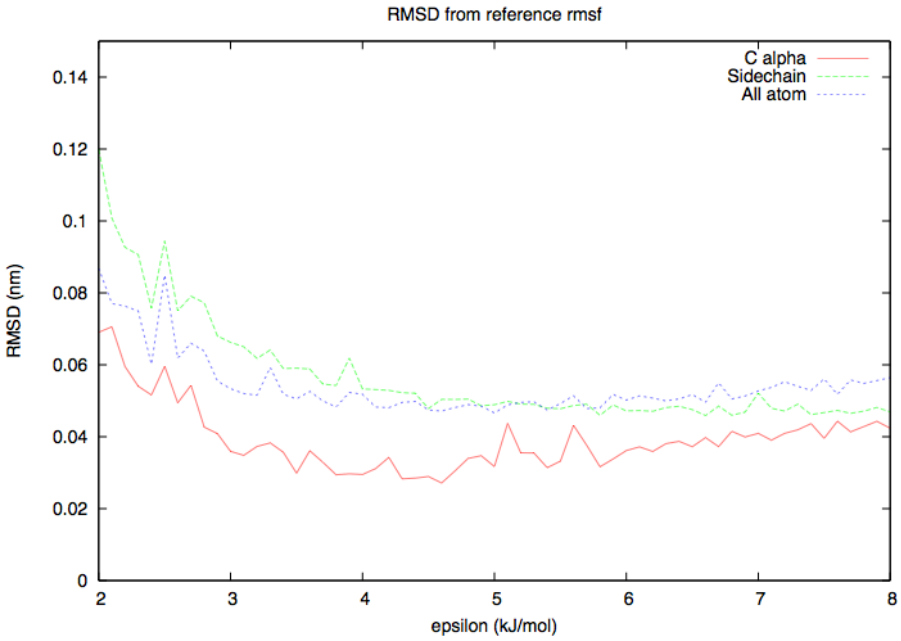

Figure 5: Root mean square deviation with respect to the reference MD of the RMS atoms fluctuations for different values of ε_{SB}.

3

FOLDING

3.1 INTRODUCTION

Protein folding is the process for which structure-based models were originally proposed. Since their first steps [57], these models fostered discussion and contributed to a deeper theoretical understanding of this subject with the help of computer simulations. The existing, remarkable body of work in this field was used to ground our model, in preparation for the treatment of conformational transitions.
I start by describing the system under consideration, then introduce the quantities useful for the analysis and finally present the results and a few additional considerations.

3.1.1 THE FOLDING BENCHMARK

The SH3 domain per se attracted interest in the study of folding processes for having its structured part mostly composed of β strands. It has become an archetype for β sheet and two-state folding thanks to many works dedicated to its characterization [70, 71]. Two perpendicular beta-sheets composed of, respectively, three and two anti-parallel β-strands build up the basic fold of SH3 domain (see fig.8). Two loops, one hairpin, one turn and a helix 3_{10} turn connect these β-strands to shape this characteristic β-barrel. We used the structure of human c-Src SH3, extracting it from structure 1FMK [72]. The simulated strand amounts to 55 residues. Its experimental transition temperature is 350 K [70].

3.2 DEFINITIONS

Protein folding inspired the definition of many dedicated characterizing quantities and the use of already existing ones from thermodynamics. For our purpose of comparing the performance of the model with known experimental facts, we chose the following ones.

AVERAGE Q For a whole protein (or a set of residues), $\langle Q \rangle$ (or $\langle q \rangle$) is defined as the average fraction of native pairs formed in the transition state at a fixed temperature with respect to the total number of native pairs. In our case, a native pair is considered formed if the separation between its atoms is within 1 Å from the native distance, and should approach 1 in the folded state and 0 in the unfolded state.
The quantity $\langle Q \rangle$ is capable of discriminating between the extremes of the transition of interest. Thus, it was used as order parameter for the heat capacity (see next paragraph) and as free energy coordinate. As a consequence, $\langle Q \rangle$ was also used to determine the folding (or critical) temperature T_f, defined as the temperature at which the two free energy minima are equal, and equivalently the probability to find the system in the native state is equal to the probability to find it in the unfolded state.

HEAT CAPACITY AT CONSTANT VOLUME The behavior of the heat capacity in function of the temperature (I will be using $C_V(T)$, the heat capacity at constant volume) is capable of providing insights on the folding transition of a protein because, at the folding temperature, it presents a peak. This quantity provided us a second facet on the folding temperature, besides the first that was indicated by the free energy profile.
We exploit the following relation from statistical mechanics between the heat capacity at constant volume $C_V(T)$ and the energy fluctuations $\Delta E = \sqrt{\langle (H(Q) - \langle H(Q) \rangle)^2 \rangle}$:

$$C_V(T) = \frac{(\Delta E)^2}{k_B T^2} \tag{3.1}$$

where k_B is the Boltzmann constant.

COOPERATIVITY INDEX κ_2 The notion of cooperativity describes the ability of a sequence to overcome energetic frustration in the conformational search. It is known that the folding of helical domains is driven by local interactions, while sheet formation and irregular patterns are rather dominated by nonlocal interactions and thus rely on a higher degree of cooperativity for folding. Here, to account for the transition cooperativity, we used the measure

$$\kappa_2 = 2\frac{T_{peak}}{\Delta}\sqrt{k_B\, C_V(T_{peak})} \tag{3.2}$$

where $\Delta = \int^{TR} dT\, C_V(T)$ is an integral across the transition region (adapted from [73]). The maximum value for κ_2 is 1. Thus we expect a κ_2 approaching this maximum value for β folding.

3.3 METHODS

I report on the analysis of a parallel tempering MD of 40 replicas in the temperature range 350-370 K, with $\varepsilon_{SB} = 3.8$ kJ/mol. The simulation was 500 ns long with exchanges between replicas attempted every ps.
The energy distribution corresponding to each temperature is reconstructed from the trajectories via the multiple histogram method [74].

3.4 RESULTS

3.4.1 FREE ENERGY OF FOLDING

Following our considerations on $\langle Q \rangle$, the critical temperature for SH_3 with our model is found at 357 K and a barrier of 14 kJ/mol separates the two states. Fig.6 shows the free energy at this temperature (solid line). A remarkable discrepancy in the population of the two basins at temperatures close to T_f is also highlighted in the image with dashed lines.

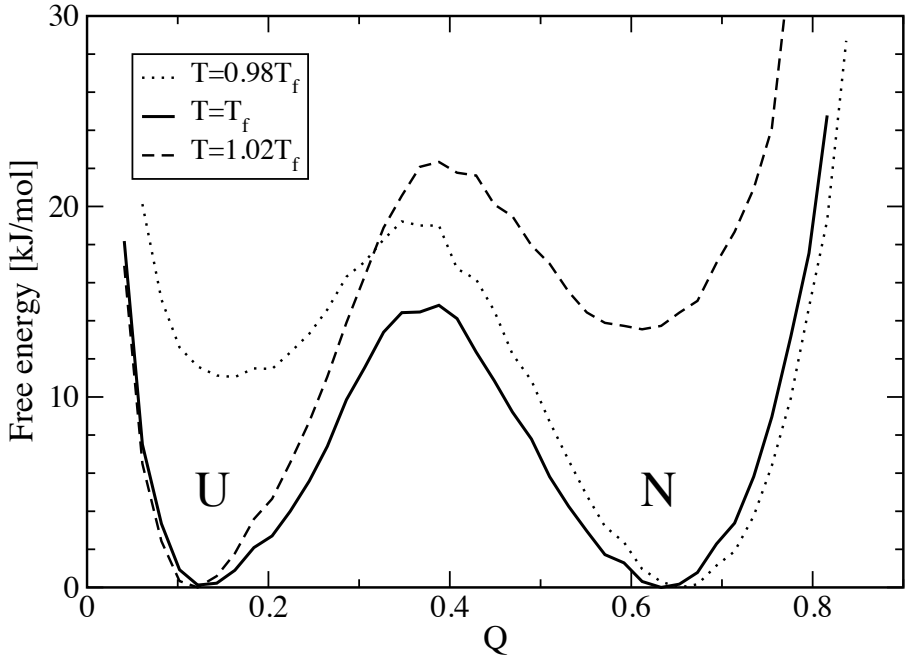

Figure 6: Folding free energy with respect to $\langle Q \rangle$ around the transition temperature $T_f = 357$ K.

3.4.2 HEAT CAPACITY

The upper panel of fig.7 shows the heat capacity at constant volume as a function of T, whose peak at 357 K coincides with the transition temperature. The fact that the transition temperature is very close to the experimental one (350 K, [70]) strongly indicates that the model blends its two force field and structure-based flavors in balanced proportions through ε_{SB}. The substantial consistency of the model's temperature to the physical temperature has primary importance here, since the force field parameters that it includes are parametrized on the standard thermodynamic definition of temperature. This is at variance with pure structure-based models, that do not contain classical force field contributions and are thus freed from the need of full consistency between simulation temperature and physical temperature. In such pure structure-based models, while the simulation temperature still drives the transition by contrasting the force of native interactions, it by no means numerically coincides with the standard definition of temperature that compares with experiments.

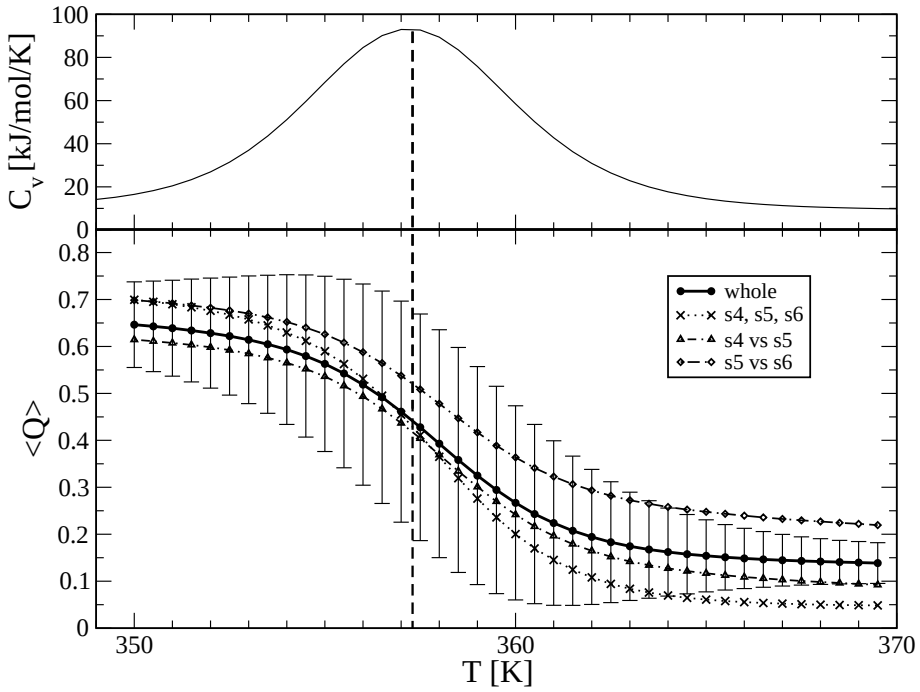

Figure 7: Heat capacity at constant volume (higher panel) and $< Q >$ VS T (lower panel) for various structural motifs.

Figure 7 also shows the evolution of different $< q >$ values with temperature. The solid line refers to $< Q >$ for the whole protein. We also calculated $\langle q \rangle$ as a function of T for the β-sheet composed by the three strands s4, s5 and s6 (see fig.8) building the main part of the hydrophobic core of SH3. This allowed to checked the degree of formation of subsets of this structure around the transition state according to our model.

The results (fig.7, lower panel) show the following local behavior:

- The whole subset of the three β strands has $\langle q \rangle$ decaying faster to low values than the whole protein, meaning that this group as a whole is unstructured in the transition state;
- The event above may be driven by the disassembly of part of the three strands. In fact, the subset of strands 5 and 6 (*s5* and *s6* in fig. 7 and 8; respectively, residues 42-47 and 50-55) reveals to be more stable than the average. This hairpin-shaped motif connected by the *distal loop* was confirmed experimentally as relatively well structured in the transition state by mutagenesis studies [75].

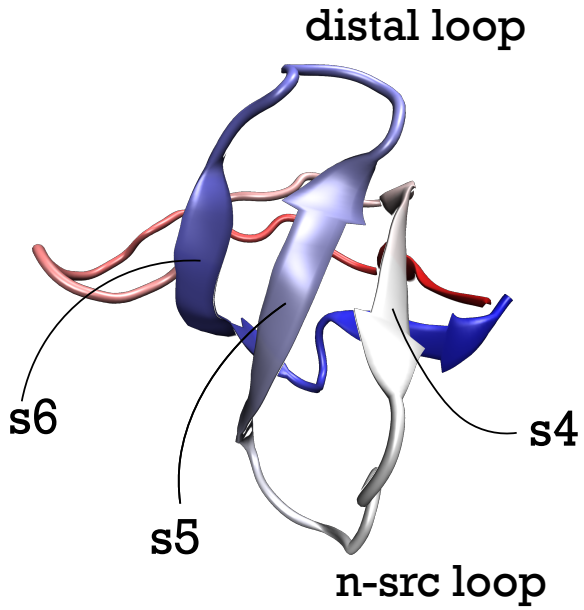

Figure 8: The SH3 domain and its characteristic three β-strands structural core.

- Strands 4 and 5, connected by the *n-src loop* (*s4* and *s5* in fig. 7 and 8), remained constantly below the average $\langle q \rangle$ of the whole protein indicating early unfolding. This notion found also confirmation in experimental mutagenesis studies [75].

The cooperativity index amounted to $\kappa_2 = 0.80$, indicating a high degree of cooperativity (maximum value for κ_2 is 1) of the transition.

Both with respect to the transition temperature and the evolution of structure loss as denaturing condition strengthen, the model is consistent with experimental indications [70, 75]. We conclude that the description of the transition state provided by the model with respect to β folding represents a positive proof on the model's performance.

OTHER VALIDATIONS A discussion of the previous results and additional characterizations (including Φ values calculation[1]) can be found at ref.[68].
In this paper, a protocol validation similar to the one described for SH3 is reported for the villin headpiece HP35 to create an analogous benchmark for α folding. The HP35 subdomain is composed of 35 residues structured in three helices. It is the extreme part of a compact f-actin binding domain located in the C-terminal part of protein villin. It is highly thermostable and has been described as the smallest monomeric sequence characterized that folds autonomously into a unique and thermostable structure, that is constituted of only naturally occurring amino acids and whose folding process does not depend on disulfide bonds or ligand binding [76]. Like SH3, its folding has been studied extensively on the computational and experimental front [76, 77, 78, 79].
I calculated the cooperativity index for HP35 for our model and, in agreement with the expectation of a lower value for α folding, it amounts to $\kappa_2 = 0.53$, which helps to confirm the model's physical reliability in combination with the $\kappa_2 = 0.80$ value for SH3 and β folding.

1 To measure the stability of a single residue in the transition, Φ values are commonly used as an indicator. In general terms, Φ values have a simple interpretation: a value approaching zero stands for a low occurrence of the residue contacts in the transition state, whereas a value close to 1 stands for a residue that participates to a highly structured strand in the transition state. Φ values can be measured both experimentally and computationally.

Thanks to the mentioned indications, we concluded the tuning phase for our model. We then proceeded to targeting it towards conformational transitions in proteins.

4

CONFORMATIONAL TRANSITIONS IN C-ABL

4.1 REMARKS ON MODELING FOLDING AND CONFORMATIONAL TRANSITIONS

Now that we have described the validation phase of the model on folding, it is worth to consider the leap we are making by transferring a mainly structure-based model to the description of conformational rearrangements. Awareness on the qualitative differences between these two transitions is necessary, as they are two close but distinct physical phenomena.
Folding is a global transition with the traits of a finite system phase transition (a peak in the derivative of the free energy with respect to an order parameter, in this case $\langle Q \rangle$, the average fraction of native contacts formed). On the contrary, the large-scale conformational transition is local since it involves a minority of the components of the system, and this fixes specific constraints. While during unfolding the system is driven out of a single minimum, a conformational transition corresponds to a controlled transition between two local minima via a transition state. Consistently, the unfolded extreme of the folding transition is high and highly degenerate in energy, while for local conformational transitions the two extremes are structurally well defined and is the region of the transition state to be the highest in free energy and entropy. Consequently, from the standpoint of modeling a temperature-induced transition (through the intensity of the native contacts ε_{SB}), in order to avoid temperature-induced unfolding of the whole protein, contacts that ensure fold stability may require a different treatment from those involved in the targeted local conformational transition.

Having said that, two-state folding has been investigated extensively on the experimental and theoretical front and provides a natural and convenient framework to start from.

***

The present chapter describes the first application of our hybrid model to a multidomain protein, a case of great interest in cancer research. I start by

describing the protein our method was applied to, c-Abl, together with the multifaceted challenges this protein presents to the scientific community and relatively to which, we submit, our model can provide insight. Next, results are presented with the discussion about how they contribute to the debate c-Abl currently fuels.

4.2 C-ABL

The term c-Abl protein indicates the cytoplasmic and nuclear nonreceptor tyrosine kinase encoded by c-abl proto-oncogene 1 (ABL1).
c-Abl kinase is expressed ubiquitously in mammals [5]. One paralogue, Arg (abl related gene, ABL2), has been found for c-Abl and with it constitutes the ABL subfamily.
Human c-Abl kinase is 1182 residues long. It is expressed in two N-terminally differing isoforms: 1b, with the N-terminal cap bound to a myristoyl fatty acid (Myr), and 1a, nineteen residues shorter and not reported to be myristoylated [80]. Next to this spliced segment, the two polypeptide chains fold identically in one Src homology 3 (SH3), one SH2 module and a tyrosine kinase domain. This linear modular arrangement of SH3, SH2 and catalytic domain is common to c-Abl and the closely related c-Src tyrosine kinase. A long C-terminal domain, named last exon region, of about 90 kDa and unique of c-Abl follows. It contains several binding and recognition sites, namely three proline-rich binding sites (PXXP) for adaptor proteins, three nuclear-localization signals, three tandemly repeated DNA-binding domains, G- and F-actin binding domains and finally a nuclear-export signal (NES)[5].
The kinase domain of c-Abl catalyzes phosphorylation of a tyrosine in peptide

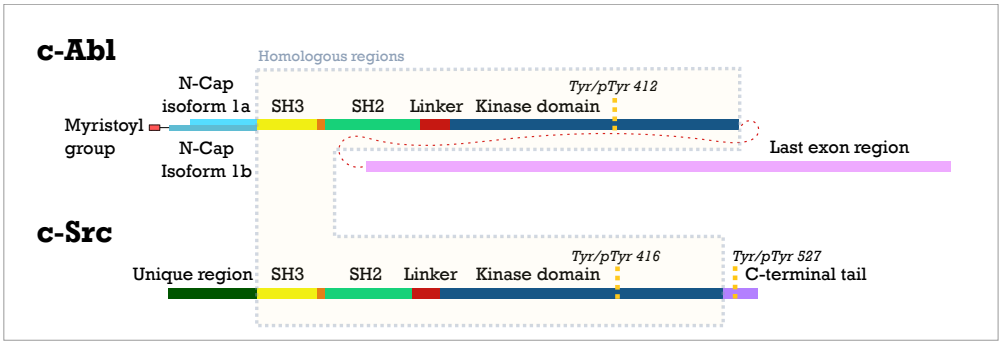

Figure 9: Modular organization of c-Abl and c-Src.

substrates. While c-Abl belongs to the ABL subfamily, and at a higher level to the group of nonreceptor tyrosine kinases, the fold of its kinase domain is a shared feature of the whole kinase superfamily, estimated to include 518 kinases [11]. It consists of a strand of about 300 residues organized in two lobes connected by a hinge region and defining a deep cleft where ATP binds. The N-terminal lobe is smaller and composed of beta sheets, except for a conserved critical helix, called αC-helix. The C-terminal lobe is mainly helical.
The catalytic domain performs the biochemical function associated to any kinase protein: the phosphorylation of protein substrates, namely the transfer of the γ phosphate of ATP to the hydroxyl groups of a serine, threonine or tyrosine residue. This is realized by a catalytic domain that exhibits the following defining traits of an active conformation.

4.2.1 General features of active and inactive protein kinase domains

activation loop A glaring, common indicator of an activated kinase domain is the position of a flexible strand of about 20-30 residues called activation loop and located between the two lobes. This region can severely contribute to inactivation by hindering substrate binding. Consistently, in an activated state it is removed from the substrate binding region (like in fig.10) and, in c-Abl, stabilized in this extended conformation by phosphorylation of Tyr 412 (human c-Abl 1b numeration[1]).

1 The numeration we will always refer to in the following for c-Abl, unless otherwise stated.

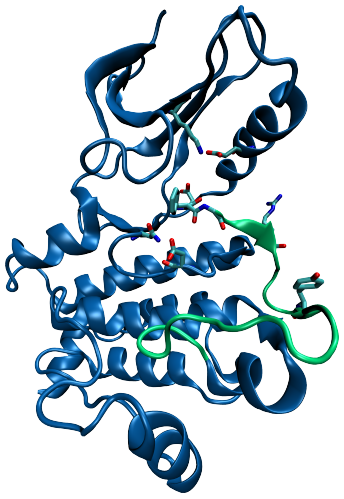

Figure 10: Kinase domain with the open activation loop in green (the comparison with opposite conformations can be better appreciated when I discuss specific conformations at page 47). Key residues of the active site are shown in licorice representation. See fig.11 for a higher detail on these residues.

ACTIVE SITE AND αC HELIX Regarding the active site, two key ionic bonds between highly conserved residue pairs have been recognized as shared among the whole kinase family [12]. One involves Asp 382, which additionally interacts with ATP, and Asn 387, needed for the binding of one Mg cation.
The state of the second salt bridge is characteristic of two main inactive conformations of the kinase domain and involves Lys 290 from sheet β3 in the N-lobe and Glu 305 from the αC helix. These two classes of inactive conformations are commonly referred to as Src-like and Abl-like inactive conformation. They differ primarily in: 1) the orientation of the sidechain of residue Glu 305 and 2) the orientation of residue Asp 400 from the DFG motif, first of a highly conserved triad (Asp 400 - Phe 401 - Gly 402) located at the base of the activation loop and at the core of the active site (discussed in the next paragraph).
Residue Glu 305 is conserved both in Src family kinases (SFKs) and c-Abl. In the active conformations, it establishes an ion pair with Lys 290 that forces it to point towards the active site and the longitudinal axis of the domain. In the main inactive conformation of SFKs this pair is disrupted: the Glu is rotated and steadily surface-exposed by another ion pair with the likely orientated Arg 405 from the activation loop, defining one feature of a Src-like inactive conformation. In c-Abl and other kinases this breakage is not observed in most of the inactivated structures, a fact considered distinctive of an Abl-like inactive structure.

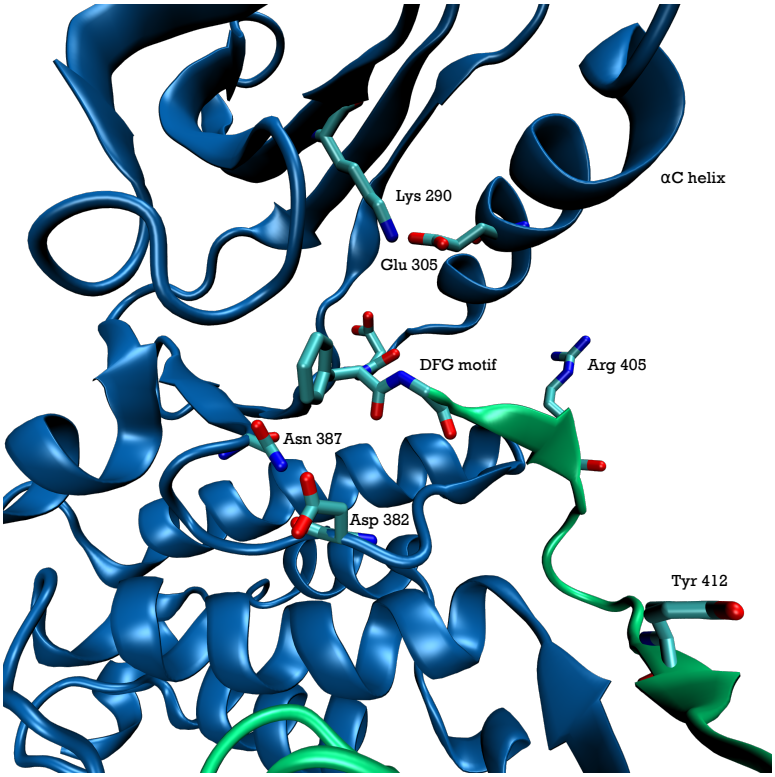

Figure 11: Detail on the active site.

DFG MOTIF A possible inactivating transition involves primarily the DFG motif, the triad (Asp 400 - Phe 401 - Gly 402), located in the active site. Asp 400 sidechain points towards the active site in the active conformation and is functional to Mg^{2+} coordination. The Asp sidechain can significantly move away from the coordination site, a fact that results in Asp 400 and Phe 401 swapping positions for steric and stereometric reasons, and that can be seen as an overall crankshaft-like rotation of the whole DFG segment (see fig. 12). The result of this movement is referred to as the DFG flipped, Asp out conformation and is sufficient for inactivation. Briefly, then, for what concerns these two critical sidechains, an active kinase domain has the αC helix Glu pointing inwards and the DFG Asp pointing towards the active site. When inactivated, the Abl-like kinase retains Glu inward-directed and has the DFG Asp pointing out, and the Src-like one features Glu outwards and retains the DFG Asp in (see table 1).
This conformational discrimination is paramount also when it comes to the development of selective ATP-competitive inhibitors. In particular, the DFG-out conformation is required for high affinity imatinib binding and provides the molecular basis for the restricted selectivity of the drug, paving the way for the groundbreaking clinical success of this inhibitor in Ph+ diseases and freezes the kinase domain in an inactive state.
Time told, though, that the association of the different conformations with inactive Src and Abl was less rigid than the expected and a Src-like inactive conformation was found for inhibitor-bound Abl [81] and vice versa [82].
Levinson et al. managed to capture in 2006 c-Abl accessing a Src-like inactive

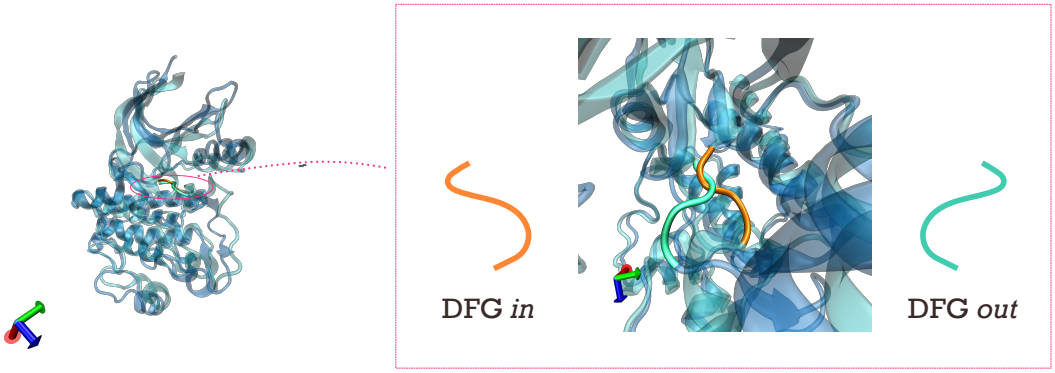

Figure 12: The crankshaft-like character of the DFG transition. See also page 47 for the orientation of Asp 400 in two experimental structures.

DFG → αC Glu ↓	in	out
in	active	Abl-like inactive
out	+ Glu 305 - Arg 405 salt bridge → Src-like inactive	intermediate

Table 1: Possible combinations of critical residues' arrangement and their relationship with activation states. Adapted from [81].

conformation (αC Glu out and DFG Asp in) in the presence of an ATP analog peptide conjugate [81]. Here, the DFG Asp 400 points towards the kinase axis and forms the mentioned critical ion pair with Lys 290. A solvent-exposed salt bridge between the αC Glu 305 and Arg 405 from the activation loop is also highlighted. This Src-like structure has been interpreted as a possible intermediate conformation towards the DFG flipped conformation and a number of correlated observations is provided on the possible flipping pathway. A special role for residue Arg 405 in the DFG flip is suggested. This residue forms part of a peculiar helix 3_{10} turn and additionally interacts with two residues from the catalytic loop, and because of this it was found to be properly oriented to catch and stabilize Asp 400 while it flips downwards and outwards from an *in* conformation. This is consistent with an intermediate minimum in c-Abl reported in [83]. Also, the αC Glu transition finds a possible electrostatic stabilizer in Arg 381 all the way to the salt bridge with Arg 405.

Complete DFG flip was reproduced in simulations of the Abl kinase domain [84, 83, 85].
In [84], pH was proposed as a factor affecting the kinetics of flip through Asp protonation and the *in* → *out* flip was detected 3 times in unbiased MD simulations in slightly diversified surroundings but always with Asp 400 protonated at 300 K and 1 bar. In equivalent conditions, the flip of unprotonated Asp did not occur. These observed flips were consistent with the counterclockwise protonated Asp-Phe passage proposed in [81]. Simultaneous to the most sterically hindered step with Phe pointing upwards, maximal displacement of the αC helix was observed.
For what concerns free energy calculations, it is necessary to sample the DFG transition up to convergence to equilibrium to estimate the free energy dif-

ference between the two DFG states. Once this level of sampling has been attained, the following formula is used:

$$\Delta F = F_1 - F_2 = -\frac{1}{\beta} ln \left(\frac{N_1}{N_2} \right) \tag{4.1}$$

where N is the number of times the given state is visited in the equilibrium ensemble. Unbiased simulations do not provide a sufficient sampling of the states, even on purpose made supercomputers, while an approach based on enhanced sampling was able to converge the free energy landscape associated with this transition [83]. This was done via parallel tempering metadynamics [52] on c-Abl and c-Src kinase domains, obtaining respectively:

- $\Delta F_{in \to out} = F_{out} - F_{in} = 4.0 \pm 0.5$ kcal/mol, with an intermediate minimum of 1 kcal/mol under the DFG-out free energy value, consistent with the indications of [81];
- $\Delta F_{in \to out} = 6.0 \pm 0.5$ kcal/mol and no intermediate minimum was found.

In [85], the comparable potential of mean force relative, obtained from umbrella sampling MD at 300 K and 1 atm, was measured in $\Delta F_{in \to out} = 1.4$ kcal/mol for c-Abl and 5.4 kcal/mol for c-Src.
A less computationally expensive scheme has been pursued in [86] and pointed out a conformational advantage of the DFG-in conformation for c-Src of 4.7 kcal/mol, and a similar but less definitive indication for phosphorylated c-Abl of 3.4 kcal/mol.
In conclusion, all computational indications, despite the usage of different methods, qualitatively agree in indicating the DFG Asp-in conformation as the most energetically favourable for c-Abl.

HINGE MOTION The catalytic activity also depends crucially on the collective, relative movement and orientation of the two lobes. They both contain the structural apparatus (like, respectively, the αC helix and the activation loop) that, with the deriving atomic critical interactions, control the catalytic cycle. Additionally, they must also open and close to allow ATP binding and ADP release, because the catalytic cleft is located between them. The fluctuations of the lobes around their juncture are thus part of the interplay of correlated interactions that compose the catalytic cycle of a kinase: they both reflect and affect

other relevant conformational transitions for protein function. This collective displacement has been described as the "breathing movement of the kinase domain" and will also be investigated in the following.

4.2.2 A MODEL FOR C-ABL REGULATION

The concept that c-Abl admits an autoinhibiting modular arrangement has found increasing support [87, 88] at the beginning of the last decade over the idea that a cellular substrate was the cause of its inhibition [5].
An important step towards this conclusion was published in 2002, when Pluk et al. showed for c-Abl 1b [87] that the whole segment C-terminal to the kinase catalytic domain (beyond residue Glu 531) is not required for the regulation of kinase activity. Instead, they showed the minimal segment required in Abl 1b to see activity suppressed includes segments 1-531, with a critical role of the strand 1-81. This was proved on c-Abl 1b, and the isoform 1a was included in the model on the basis of evidences of an analogous behavior. This implies that this strand contains all the key factors for c-Abl full catalytic control. We know this includes one C-terminal unstructured tail, one SH3 domain, one SH2 domain, a linker to the kinase domain and the kinase domain itself.
The SH2-kinase domain motif is known to specifically bind phosphotyrosines (pTyr) [89, 90] and is highly conserved in all cytoplasmic tyrosine kinases [91]. In this scenario SH3 may respond to more specific needs for c-Abl regulation rather than SH2. SH2 comprises about one hundred residues folding in a central anti-parallel β-sheet core surrounded by two α-helices [92], and pTyr binds to a positively charged pocket on one side of the β-sheets. Like the kinase fold, also the SH2 fold is more conserved than its sequence, a fact that realizes not only the primary requirement of specifically binding pTyr residues, but also sequence specificity, that is the capacity of selecting the binding partners by their ability to bind the tyrosine-surrounding residues (this ability has been reported on sets of residues that can span, assuming pTyr at position 0, the range (-6,+6), but most often (-2, +4) [93]).
The abundance of crystal structures published in the last decade contributed increasing evidences to support a model for c-Abl autoinhibition originally introduced in [94, 95]. We will now describe this model with the aid of a few crystal structures that also represented an important experimental underpinning for the original work illustrated in this thesis. Especially important for

this description are two structures deriving from the same asymmetric unit that contains two molecules at 3.4 Å resolution (structure B of [88], 1OPL) in two dramatically different states of assembly (similar to those shown in fig.13). In the first molecule of 1OPL, a major portion (residues 81-531) of the strand capable of autoinhibition (1-531) of c-Abl is revealed. Its arrangement consists in a compact organization of the three modules, with the SH3 and SH2 domains docked distally to the active site and respectively adjacent to its N- and C-lobe. The linker connecting SH2 and the N-lobe runs throughout the whole length of the docking site, mediating SH3 anchoring via intramolecular interactions with the N-lobe, while allowing direct interaction between SH2 and the C-lobe. In this modality, as well as in any other else ever observed in this protein, the modules never get to occupy the region involved in substrate binding and activation loop movement, but rather the opposite side of the kinase domain, often referred to as the backside region. Such disposition rigidifies the hinge motion and this effect echoes on other relevant conformational transitions. In fact, when the regulatory apparatus is docked parallel to the kinase domain in the autoinhibited conformation and this motion is constricted, so are those of the activation loop and of the DFG motif because of their location between the two lobes.

The interaction between SH3 and linker in the assembled arrangement is also a contributing factor in the inactivation process. Its robustness relies on the conserved property of SH3 of specifically binding to polyproline motifs with the consensus PXXP and forming a left-handed type II helix [96]. In consideration of this notion, two prolines in contact with SH3 stand out in the linker's three dimensional arrangement. In fact, their mutation was an important part of the engineering process of a constitutively activated construct that led to one of the rare structural traces of activated c-Abl [97]. The strength of c-Abl linker/SH3 interaction was shown to conserved even in absence of the kinase domain [98]. A structure at higher resolution that shows high similarity to the mentioned assembled structure is presented at 1.8 Å resolution in the same paper for the murine c-Abl 46-534 segment[2], to which an inhibitor and myristic acid were added, respectively, during purification and crystallization. It was the first time that a similar multidomain structural arrangement was revealed in c-Abl, while an analogous, in spite not identical one, had already been described for c-

2 Murine c-Abl differs from human protein in residue 336, where is respectively a Ser and an Asn.

Src [99] and Hck [100]. The intensity of the direct interaction between the SH2 and C-lobe interface has been experimentally proven in hydrogen exchange mass spectrometry (HX MS), which revealed this interface remained protected by deuterium exchange for longer than any other region of the construct. Also, myristoyl group removal caused disruption of this interface and the deuterium exchange levels show consistency with the establishment of the top SH2-N-lobe interface [101].

These arguments all support, hence, the need of a more clear view on the machinery of c-Abl and of the closely related c-Src, that shares much of the auxiliary setup. In particular, understanding in which way the same adaptor domains may influence dynamics and contribute to explain opposite binding affinities to inhibitors is a challenging question of clinical relevance. The remarkable dimension of this system, about 500 residues and 60 kDa, puts it at the limit of current capabilities both on the computational front and on the structure-biological one. This fact motivates the relative rarity of indications regarding this system with respect to those regarding the isolated kinase domains.

The second 1OPL molecule depicts the stretch 140-238 and 252-518 of human c-Abl 1b at 3.4 Å resolution (second molecule of structure B, 1OPL [88]). It represents a structure of c-Abl in which only SH2 and the kinase domain are resolved in a peculiar arrangement. This arrangement, defined unusual by the authors in 2003, represents SH2 located on the top of an activated kinase domain with the open activation loop. The protein was found intact and completely myristoylated by mass spectrometric analysis.

Comparison of these opposite arrangements also reveals a conformational transition in the terminal helical region of the C-lobe, helix αI. In the context of the myristoyl-free kinase domain alone, this helix has sometimes been revealed in a distinguishable array of about six turns [81, 102]. In the assembled organization and in the myristoylated kinase domain alone, the number of helical turns breaks down to 4 (αI) + 3 turns (αI′): the helix is bent of about 90° towards the N-lobe, where it engages in interactions through Ile 521, Val 525 and Leu 529 with the myristoyl group. Such an αI-I′ arrangement is the only one compatible with SH2 docked at the base of the C-lobe. In the mentioned activated layout with SH2 anchored on the top interface, the αI helix conformation is compatible with the extended arrangement, but is one of the many structures whose αI helix is not fully resolved. The findings of [88] support a view for the αI helix

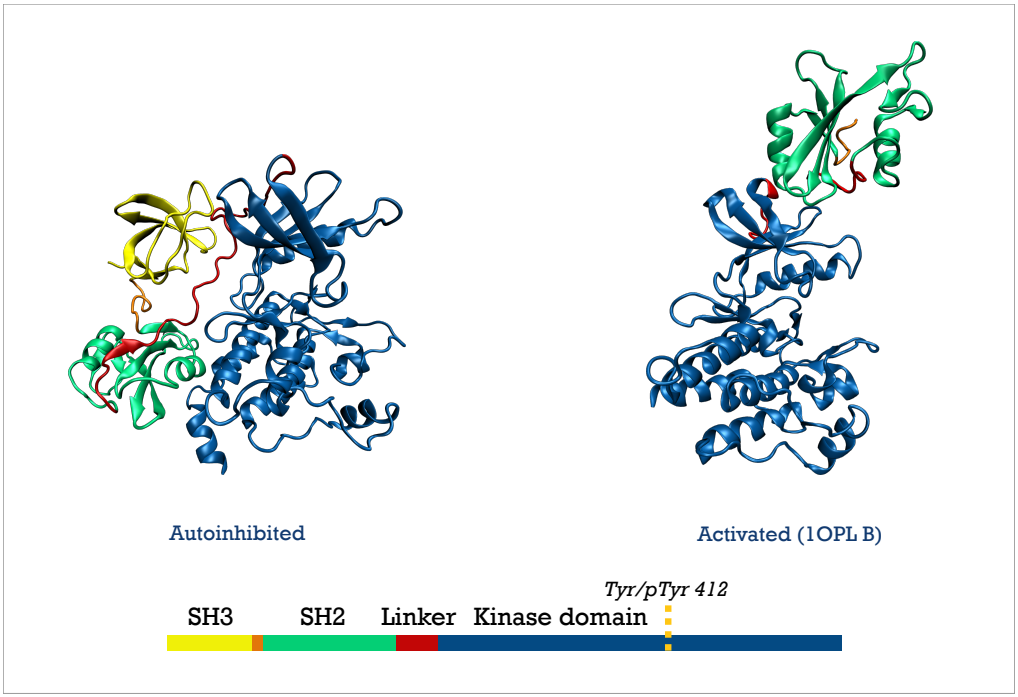

Figure 13: Comparison of c-Abl in the autoinhibited and activated configurations.

bending as a well-defined transition induced by myristate binding, intrinsical of the kinase domain and precursory of SH2 binding, rather than triggered by it.

It is interesting to note here that, in c-Src, this C-terminal helix is replaced by a tail whose phosphorylation induces SH2 binding on the lower lobe. Moreover, the oncogene product v-Src lacks the C-terminal tail but, unlike c-Abl, shows unregulated catalytic activity. This is indicative of the dependence of c-Abl on a robust autoinhibitory alternative and of how essential the phosphorylated tail is to regulation in c-Src.

After the findings described in [88], the study of the modules' extended layout was further pursued in search of a possible functional value via small angle X-ray scattering (SAXS) [97]. The resolved open loop made 1OPL B a candidate for a possible hallmark arrangement of the activated state. To test this hypothesis, a mutated form was engineered so to silence those segments involved in known inactivating interactions: starting at SH3, the protein had the autoinhibitory potential of its N-cap (residues 1-82) and myristoylation site removed,

and two SH3 binding prolines from the SH2-kinase linker were mutated in glutamic acids so to disrupt the linker-mediated packing of SH3 (P242E, P249E) to the N-lobe, typical of the autoinhibited form. Kinase assays confirmed an activity 14 times higher in this construct than in the wild type [97, 103]. The kinase domain and SH2 arrangement, as represented by the SAXS shape reconstruction [97], shows high consistency with the putatively active crystal structure. The novel top interface covers 1100 Å^2 and, on the side of SH2, it involves: the C-terminal half of helix αA, strand βG, and three loops. The N-lobe surface engages at the interface via strand β1, the loops between strand β3 and helix αC, and strands β4 and β5. The core interaction involves Ile 164 of the SH2 domain, which is buried at the interface and interacts with Thr 291 and Tyr 331 of the kinase domain.
The role of Ile 164 has then been further investigated by mutagenesis. Targeted mutation in a glutamate resulted in drastic reduction of kinase activity [93] that mechanistically illuminates the critical ability of Ile 164 of maintaining the SH2 - N-lobe interface suitably arranged for catalysis. This fact also indicates a twofold regulatory character of this domain, whose two distinct interfaces act by respectively preventing and promoting kinase activity, an effect for which the presence of SH2 has been found necessary and sufficient [93]. The SAXS shape reconstruction for the construct SH3-SH2-kinase domain also advanced knowledge in relation to two elements that were previously unresolved: SH3 matched with a density region on top of SH2 and a second density region adjacent to the N-lobe and SH2 is compatible with the linker's volume.
The work of [97] also presents 2FO0, an assembled X-ray structure of human c-Abl at 2.27 Å resolution and entailing residues 1-531 (with residues 15-56 deleted), that has been a fundamental reference for the work described here.

We described a scenario where the SH3 and SH2 domains and their atomic contacts with the interfaces of the catalytic domain are causally linked to the activation process of the two proteins. The details of the role of these domains in regulation and the elucidation of the exact differences between c-Abl and c-Src is a subject of active research. It is clear, in general, that the design of a healthy protein encodes a precise balance of populations in its complete three dimensional structure. The case of c-Abl, and in a certain measure also of c-Src, is one of a strong asymmetry towards a basin of tightly down-regulated conformations.

To investigate this, we devised a computational scheme capable of reproducing in silico several large-scale conformational transitions in one single trajectory and applied it to the construct of SH3 + SH2 + kinase domain of c-Abl, whose dysregulation can have oncogenic effects and is strongly related to the auxiliary domains.

EXPERIMENT DESIGN "What is the energetic effect of the presence of the two regulatory domains on the transitions of the kinase domain?" This is the sort of question that our model is the optimal tool to address, rather than recovering an absolute energetic information on a specific system. Classical MD simulations are still irreplaceable for this latter task.
In pursue of an answer to this question, a comparative scheme has been established. We designed a multistate topology tailored on the c-Abl transitions described above. Then, with this topology, we simulated three constructs of increasing size: one including the kinase domain only, a second composed by the kinase domain and the SH2 domain, and a third built as the kinase domain + SH2 + SH3 domains. The same free energy profiles were calculated for the three constructs with the final goal of comparing them and isolating the role on the targeted transitions of the auxiliary domains and the correlation, possibly allosteric, of distant regions in or outside the kinase domain.

4.3 METHODS

4.3.1 STRUCTURE-BASED TOPOLOGY

PDB STRUCTURES CHOSEN A requirement to include a transition in the topology was that its extremes were experimentally determined and deposited in the PDB data bank.
The pdb structures and the transitions encoded in the Hamiltonian as native states after contact extraction and merging (see below the paragraph "Building the multistate topology") are shown in the following images and described in their most relevant characteristics.

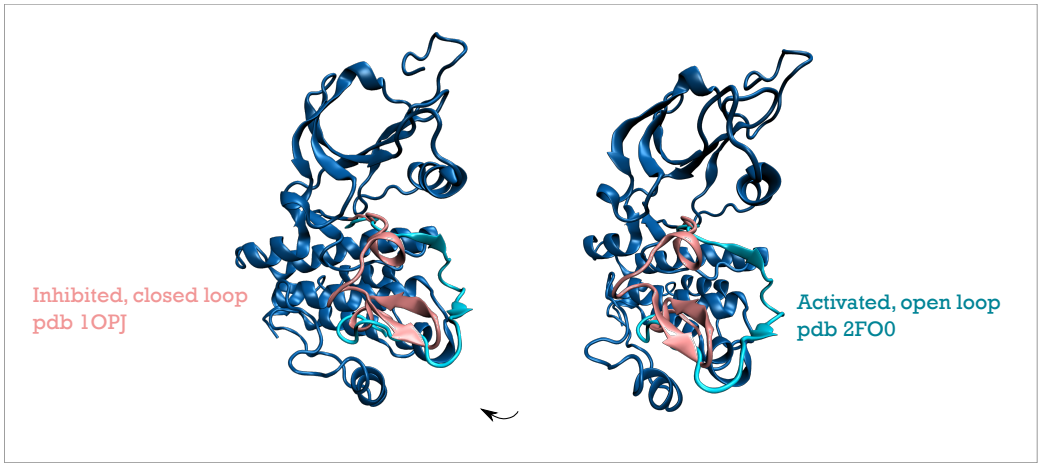

Figure 14: **Transition 1: the activation loop opening.** The loop is open in active kinases (representative pdb: 2FO0), closed in inactive (representative pdb: 1OPJ). Here the conformations are shown from two slightly different angles.

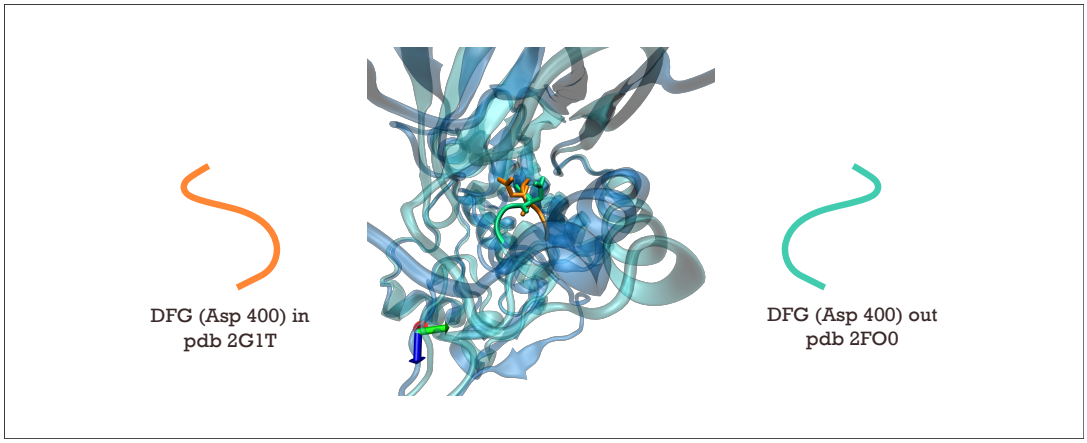

Figure 15: **Transition 2: the DFG flip.** Asp 400 *out* in inactive c-Abl (representative pdb: 2FO0), *in* in the active kinase (representative pdb: 2G1T).

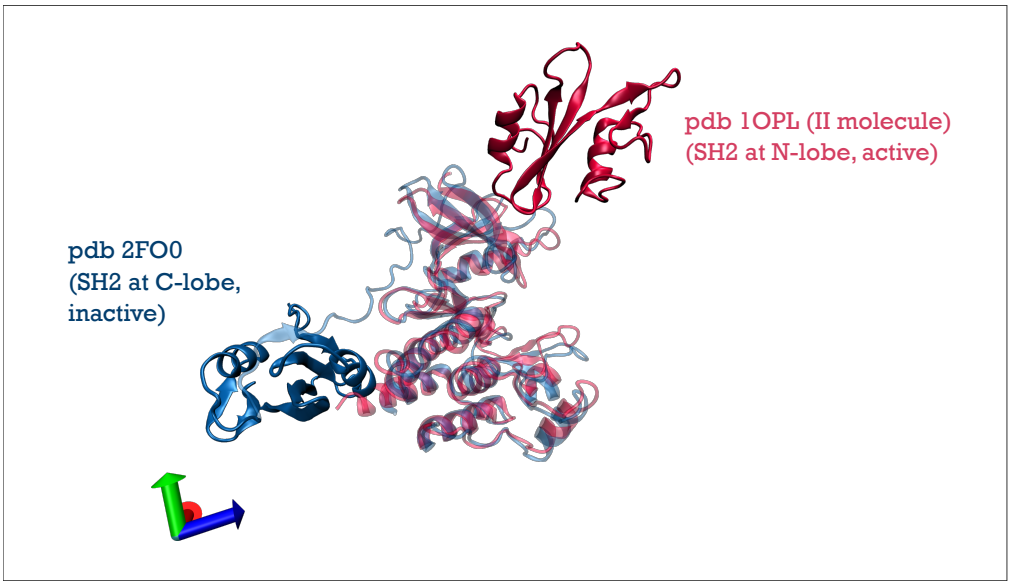

Figure 16: **Transition 3: SH2 position.** In the autoinhibited 2FOo structure, SH2 is docked to the C-lobe. In a fully active conformation (1OPL, second molecule, phosphorylated at Tyr 412), SH2 has been revealed adjacent to the N-lobe.
Particular care was given to this structure before inclusion in the potential energy function. It derived from an equilibration of 1OPL for 210 ns in explicit water performed in the research group, which ensured the removal of possible crystallographic artifacts and the inclusion of an accurate local energy minimum in the topology. Further equilibration after the addition of the missing linker and other automated structure optimizations before contacts' merging was performed for full consistency with the experiment design. The relative native contacts account then for an active state with respect to the activation loop (phosphorylated) and the SH2 arrangement, thus broadening the spectrum of our topology and analysis from complete autoinhibition to full activation. The atomic contacts' numbering of this structure was mapped to the one of the simulated unphosphorylated structure before merging the contact lists.

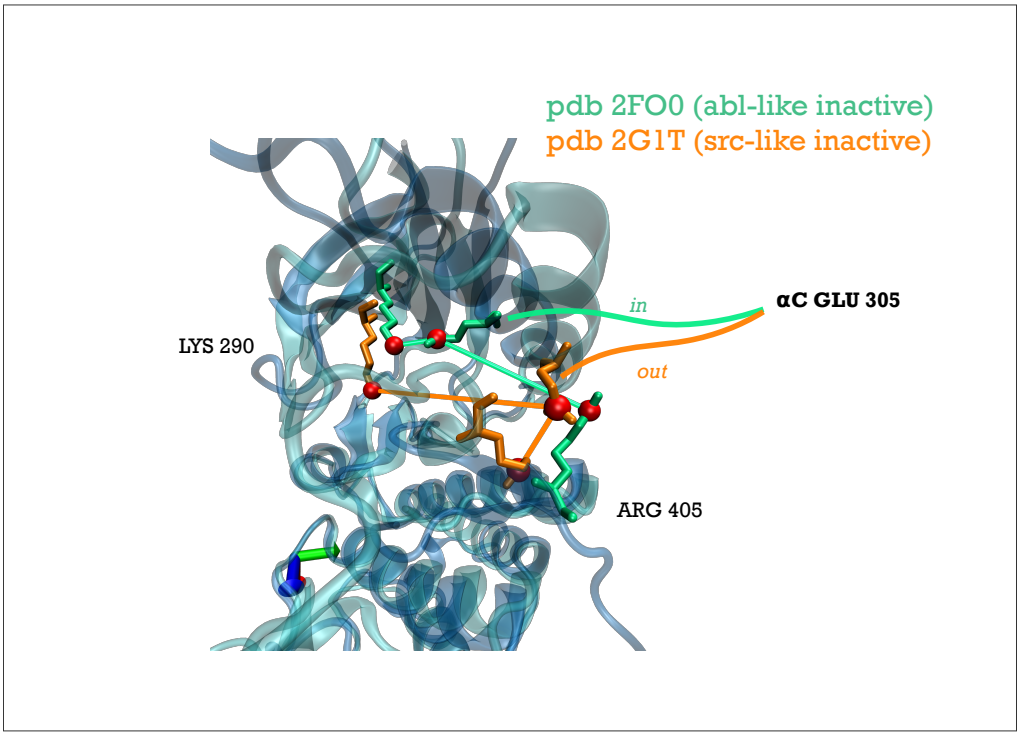

Figure 17: **Transition 4: αC-Glu orientation.** Glu 305 *in* in active kinases (representative pdb: 2FO0), out in inactive (representative pdb: 2G1T). More in general the structure 2G1T shows the feature of a Src-like inactive conformation, with disruption of the Lys 290 - Glu 305 ion pair in the active site and establishment of the solvent-exposed pair Glu 305 (αC) - Arg 405 (activation loop).

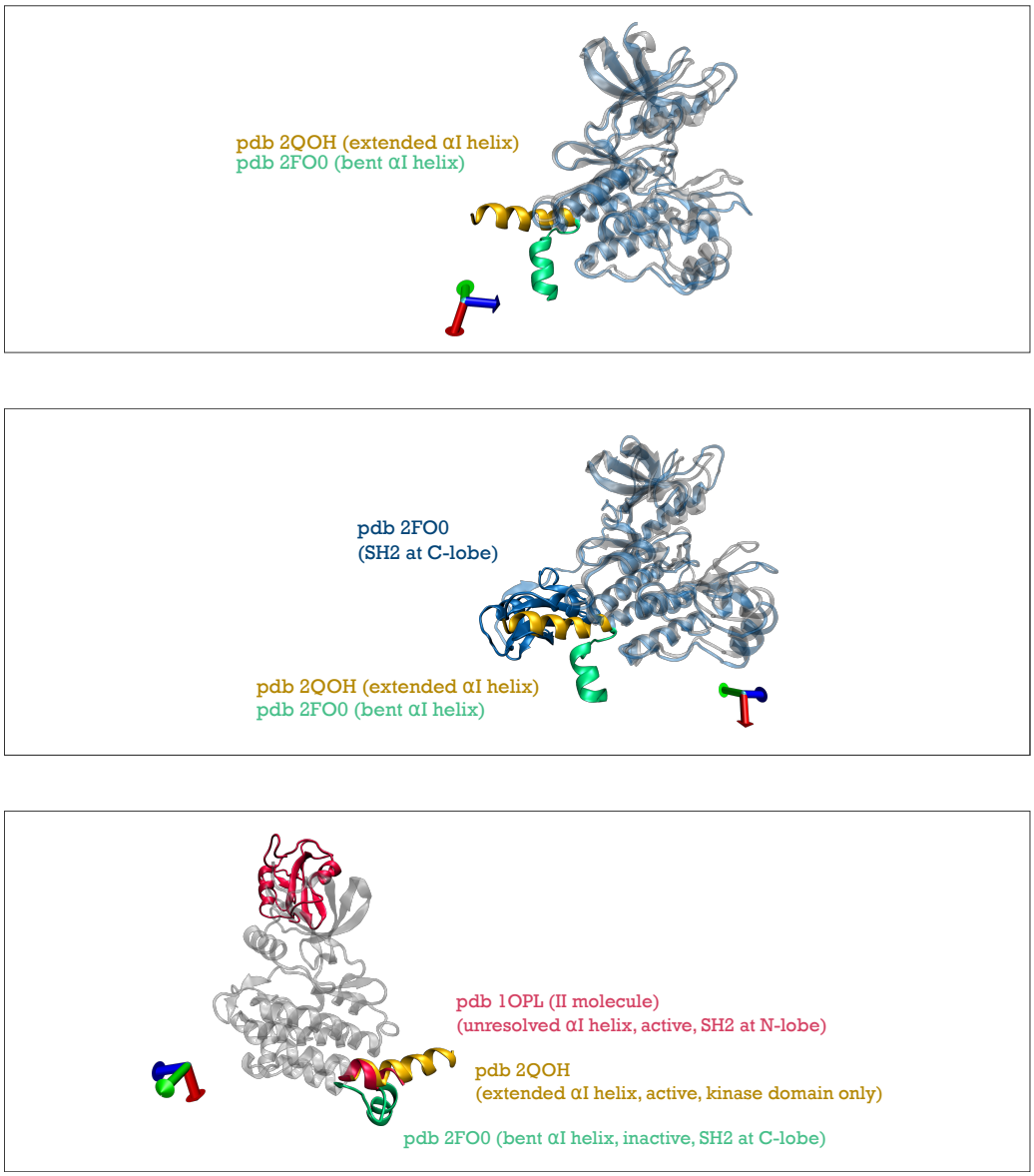

Figure 18: **Transition 5: αI helix bending.** The bending of the αI helix is associated to myristate binding. The active structure 2QOH exhibits the extended conformation. In 2FO0 (inactive) this helix is bent due to the presence of the myristate. Superposition with 2QOH shows that the relative SH2 location is not compatible with the extended layout. Additionally, we can observe that, although in the fully active 1OPL this helix is not fully resolved, its shape is compatible with the active 2QOH, rather than with the bent helix from 2FO0 (inactive).

CONTACT PAIRS EXTRACTION The chosen structures were downloaded from the Protein Data Bank and validated with MolProbity [104]. Then, maximal sequence consistency was enforced on them: eventual point mutations (none of which of clinical relevance, but rather declared either as engineered mutations or due to expression from a species different from Homo Sapiens) were revealed and adjusted towards the human c-Abl sequence. Each structure was then equilibrated in explicit water (TIP 3-point model) for at least 5 ns at 310 K with AMBER99SB-ILDN* force field [105, 106, 33, 107]. Two structures contained a phosphorylated residue (2FO0 a serine, located in a segment not included in this discussion, while 1OPL contained a tyrosine): I integrated the required parameters for these residues in the force field in use, deriving them from [108]. After equilibration, the structures were minimized in vacuum. At this point, a first list of pairs in contact was extracted according to the shadow approach explained at page 19, with cutoffs of 5 Å and 35°, and a threshold of at least three residues of separation in the sequence. Spurious contacts due to transient fluctuations in these lists were filtered out by retaining only those contact pairs for which the quantity $d + \langle d \rangle < 5$ Å, where d is their average separation distance and $\langle d \rangle$ its standard deviation during the last 2 ns of equilibration. Many public computational tools helped to perform this process: VMD [109], Pymol [110], SWISS-MODEL and Swiss-PdbViewer [111]. Table 2 contains the specifics of the five simulated experimental structures chosen in this phase.

BUILDING THE MULTISTATE TOPOLOGY The structure-based part of the topology is implemented by a Lennard-Jones interaction for each native interaction. The list of the native interaction results from merging the five lists of native contacts. Table 3 focuses on the three target structures and on the ultimate list of native contacts used on each system. The LJ minimum is set to 4.5 kJ/mol per native contact, except for 328 contacts of the activation loop deriving from non-closed conformations, which were rescaled to 2.25 kJ/mol because escape from this conformation resulted particularly difficult for the systems. This measure follows from the considerations at §4.1 and, given the comparative nature of the study, its impact is considered acceptable. The distance of the LJ minimum for each pair is the average distance calculated on 2 ns of equilibration in explicit water. When one pair pertained to more than one native structure, the shortest distance was chosen for the Lennard-Jones function minimum. The

Structure (pdb code)	Simulated strand	Number of atoms (including H)	Number of native pairs extracted
1 (2FO0)	65-530 (466 residues, SH3, SH2, linker, kd)	7379	2752
2 (1OPJ)	248-529 (282 res., kd)	4605	1747
3 (1OPL)	140-237, 251-521 (368 res., SH2 and kd)	5900	2298
4 (2QOH)	251-530 (278 res., kd)	4509	1733
5 (2G1T)	251-526 (276 res., kd)	4446	1721

Table 2: Extraction of the native pairs. The table shows quantitative details regarding the treatment of the five experimental structures underpinning the multistate topology. Under the total number of residues, the names of the domains entailed by each strand are reported.

Structure	Simulated strand	Number of atoms (heavy atoms)	Number of native pairs included in the topology
kinase domain	260-530 (271 res.)	2193	3427
kd + SH2	140-530 (391 res.)	3158	4585
kd + SH2 + SH3	81-530 (450 res.)	3620	4969

Table 3: Quantitative details on the structures simulated with the multistate topology.

Calculation	Simulation time ($\star$: approximated)	Average replica acceptance rate
kinase domain	1 μs	11.6 %
kd + SH2	772 ns*	8.1 %
kd + SH2 + SH3	860 ns*	6.7 %

Table 4: State of advancement and acceptance rates PTmetaD calculations.

classical part of the topology was built with AMBER99SB-ILDN* force field. According to one of the versions of the model developed in the research group, the parameters that refer to the dihedral angles of the backbone normally referred to as Φ^1 ($N^1 - C^1_\alpha - C^1 - N^2$) and Ψ^1 ($C^1 - N^2 - C^2_\alpha - N^2$ are not from the force field but are fitted parameters that comply with the notorious Ramachandran plot [112] for angles $[\Phi^1, \Psi^1]$, where the indexes 1 and 2 indicate two adjacent amino acids.

SIMULATION RUNS We ran a parallel tempering metadynamics (PTmetaD [52]) on six temperatures (range: 300-350 K, with a stride of 10 K) on each of the three systems evolving under the action of a stochastic dynamics, with friction coefficient proportional to the particle mass. Simulations were realized with GROMACS 4.5.5 and PLUMED 1.3, with replica exchanges attempted every ps. Table 4 compares the specifics of the three calculations with a focus on the replica exchange performance.

4.3.2 METADYNAMICS SETUP

GENERAL PARAMETERS Gaussian hills of height 2.0 kJ/mol for the kd only and 5.0 kJ/mol for the bulkier systems were deposited every 4 ps. The well-tempered bias factor was 5.0.

COLLECTIVE VARIABLES The collective variables are defined as:

CV1) S-path on 3 frames for the aloop transition. The references for the frames, calculated on 474 atom pairs from backbone and sidechains directly engaged in the transition, were: frame 1) 2FO0 (open loop); frame 2) 2G1T (semi-closed); frame 3) 1OPJ (closed loop). The gaussian width on this dimension was 0.03 units.

CV2) S-path on 2 frames for the DFG transition on 108 atom pairs. The references for the frames were: frame 1) 2FO0 (DFG Asp in); frame 2) 1OPJ (DFG Asp out). The gaussian width was 0.02 units.

CV3) Distance between the centers of mass of the groups of atoms at the interface between SH2 (27 atoms) and the C-lobe of the kinase domain (36 atoms) in the inactive conformation. The gaussian width was 0.12 nm.

CV4) Same distance as above, but for the active conformation: distance between the centers of mass of 18 atoms at SH2 interface and at the N-lobe of the kinase domain (22 atoms). The gaussian width was 0.12 nm.

A POSTERIORI VARIABLE EXTRACTION AND REWEIGHT Given the quantity of conformational information contained in each simulation, the analysis was deepened with the extraction of other relevant quantities for the description of processes of interest. After extraction from the biased metaD simulation, the associated canonical distribution was reconstructed as described in [54] through its implementation from the PLUMED package. The extracted and reweighted observables are:

- *Reweighted distance 1: Ile 312 CG1 - Asp 400 CG.* We found that the second CV, the s-path of two frames for the DFG transition, was efficient in coaxing the DFG flip but not as capable of distinguishing between the two extremes as a free energy profile coordinate. This motivated the extraction of a more linear quantity: the distance between Ile 312 CG1 in the αC helix and Asp 400 CG in the DFG motif (short when DFG Asp is pointing outwards, long when DFG is pointing inwards.).

- *Reweighted distance 2: Lys 290 NZ - Glu 305 CD.* The ion pair between Lys 290 NZ and Glu 305 CD in the active site is critical for activity, as said at §4.2.1. When the bond is formed, the kinase may be active or Src-like inactive; when it's not, the conformation is compatible with the Abl-like inactive.

- *Reweighted distance 3: Glu 305 CD - Arg 405 CZ.* This quantity accounts for the formation of the external salt bridge that engages the turning glutamic acid from the αC helix in the Src-like inactive conformation. It is minimal in such condition, and higher in any other one. See page 39 and figure 17.

- *Reweighted distance 4: Val 299 CA - Tyr 432 CA.* To measure the hinge motion (see page 41) by means of a distance, the choice fell on two amino acids that do not participate to local transitions and stay embedded in the lobe they belong to. These amino acids are Val 299 CA for the N-lobe and Tyr 432 CA for the C-lobe.

Table 5 contains the values of these observables calculated for the five native structures for numerical comparisons. The same values are also recalled in the free energy plots. Useful information to read the plots are summarized in table 6.

Structure (pdb code)	CV1 [a.u. ($\sim$ frame index)]	CV2 [a.u. ($\sim$ f. i.)]	CV3 [Å]	CV4 [Å]
1 (2FO0)	1.095	1.006	2.6	47.8
2 (1OPJ)	2.902	1.994	/	/
3 (1OPL)	1.176	1.184	49.5	3.6
4 (2QOH)	1.374	1.280	/	/
5 (2G1T)	2.004	1.202	/	/
	reweighted distance 1 [Å]	reweighted distance 2 [Å]	reweighted distance 3 [Å]	reweighted distance 4 [Å]
	6.1	3.2	12.0	32.8
	10.3	3.7	9.8	29.1
	11.0	3.2	12.6	29.3
	12.4	3.7	14.0	28.5
	12.0	15.6	4.4	29.2

Table 5: Values of the direct and extracted CVs associated to the five native structures contributing to the structure-based topology.

Name	Description	Protein region	Trend (min $\rightarrow$ max value)
CV1	s-path on aloop (3 frames)	Activation loop	open $\rightarrow$ closed
CV2	s-path on DFG (2 frames)	DFG motif	out $\rightarrow$ in
CV3	distance from SH2 lower interface	SH2 and SH3 position	inactive $\rightarrow$ active
CV4	distance from SH2 higher interface	„	active $\rightarrow$ inactive
reweighted distance 1	distance Ile 312 CG1 Asp 400 CG	DFG	out $\rightarrow$ in
reweighted distance 2	distance Lys 290 NZ Glu 305 CD	active site	active $\rightarrow$ inactive (Abl-like)
reweighted distance 3	distance Glu 305 CD Arg 405 CZ	external salt bridge Src-like inactive	inactive (Src-like) $\rightarrow$ any other state
reweighted distance 4	distance Val299 CA Tyr432 CA	hinge motion	

Table 6: Key features of the direct and extracted CVs.

4.3.3 ADDRESSING CONVERGENCE

The free energy space associated to these calculations is remarkably large, especially when the auxiliary domains are included and with them four collective variables are involved. Hence, quantifying the extent to which this highly-dimensional space has been explored may represent a useful orientation tool before analyzing the results. In the case of the kinase domain, where the CVs were two, we claim convergence was attained by observing that the hills' height goes smoothly to zero (see figure 19), in agreement with the well-tempered metaD scheme. We divided the free energy space in elements of volume of the size of ten times the gaussian width and counted how many of these hyper-cubes where visited since 600 ns until the end of the simulation at $1\mu s$: the visited volume results 97% of the total.

The matter changes for the systems including the auxiliary domains. In this case, the free energy space has two more dimensions, so two orders of magnitude more of hyper-cubes to visit. The same quantity, the percentage of volume visited by the system, drops to 36% for the SH2 + kinase domain and 31% for the SH3 + SH2 + kinase domain. The hills' height also reflects some distance from the condition of full exploration of the free energy space. Yet, we can add to this picture that the main part of the non-explored space is arguably ascribable to the large portion of the solid angle allowed by the full extension of the linker to the detached group of the auxiliary domains. For our purposes, the free energy profile in the dimension of CV3 and CV4 features a large entropic minimum that is as highly degenerate in free energy as large to fully explore. This minimum is already present in the free energy profiles we present, together with the other two conformations of interest, with SH2 docked at one of the two interfaces. Thus, all the conformations we were interested to correlate are present in the trajectories we present and discuss here. Completion of a uniform sampling of the free energy space will possibly confirm or provide quantitative adjustments to the considerations these simulations already allow and that we now proceed to describe.

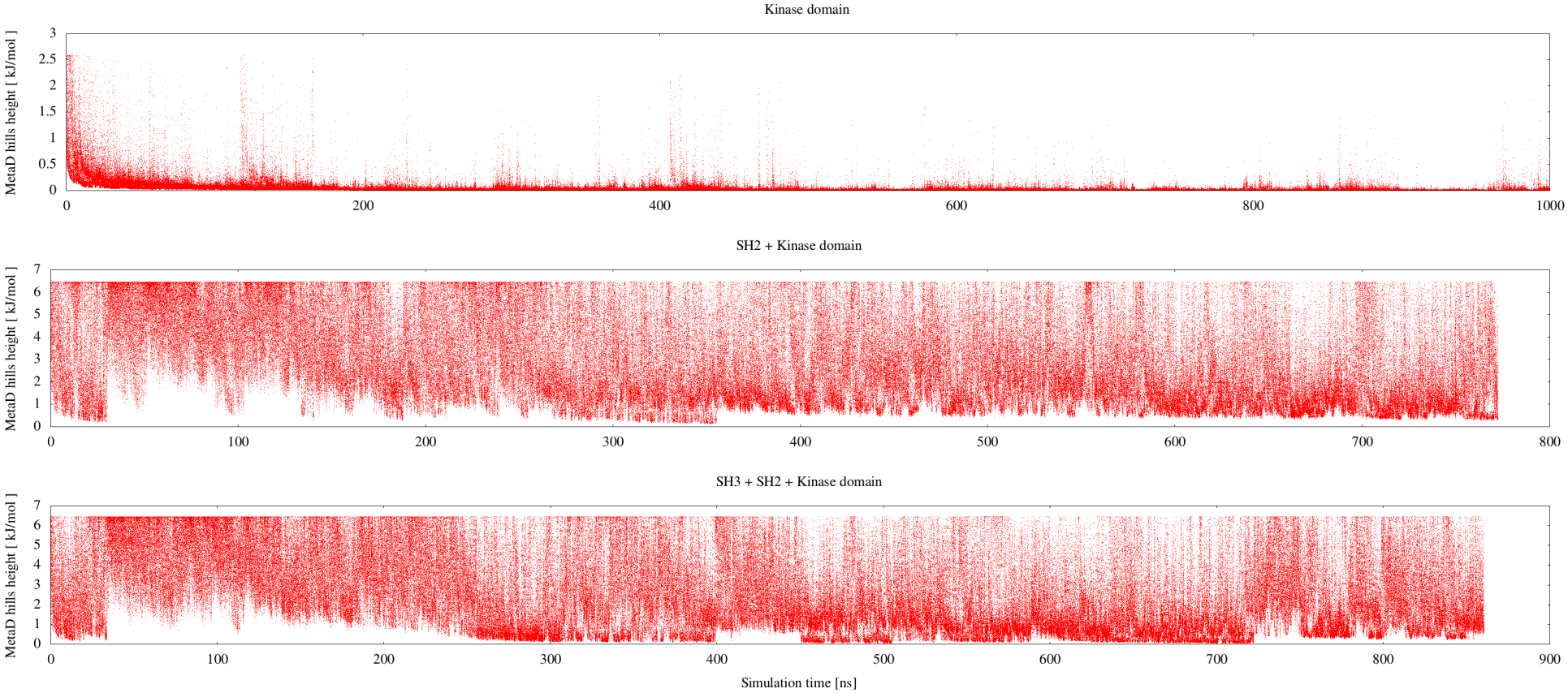

Figure 19: Comparison of the metaD gaussians' height of the simulation at 310 K for the three systems.

4.4 RESULTS AND DISCUSSION

4.4.1 EVOLUTION OF THE TRAJECTORIES AT 310 K

This first block of results responds to a descriptive intention. Here I describe the details of the evolution of the three trajectories that compose this study by looking at each CV separately, for each trajectory. The plots under discussion are collected in the appendix, starting from p.79. The aim is helping the visualization of the actual sequence of conformational events that the systems undergo that are the basis of the discussion contained in the next sections. The simulation of the only kinase domain obviously lacks the CV that refer to the auxiliary gear.

The aim of adding two CVs triggering SH2 detachment and attachment to the two interfaces was elucidating the physical role of the auxiliary domains in activation and downregulation. Although, this came at the price of enlarging by two orders of magnitude the configurational space to sample, with the consequences described at §4.3.3. We remark that, in both simulations including the auxiliary domains, the goal of modeling the conformational transition between the two extremes has been reached, at least qualitatively. In fact, if we observe the plots that refer to CV3 and CV4 of the simulation of SH2 + kinase domain (fig. 35 and 37), the auxiliary module samples repeatedly the N-lobe interface, while in the homologous case of the same CVs for SH3 + SH2 + kinase domain (fig. 36 and 38) this happens with much more rarity. This fact is definitely caused by the presence of SH3 and its interactions with the linker that result in a consistent hindrance to the full extension of the linker and the attached domains. We will give to this occurrence appropriate space in the next paragraph, as it reflects a constitutive characteristic of the full-length protein with a heavy impact on the overall dynamics of the system.

Having anticipated this with the aim of making the interpretation of the results easier, we present here the evolution of each of the individual variables during the course of the simulation. The upper panel plots the CV values at each step of the simulation and give a graphical intuition of its evolution. The lower panel depicts the relative free energy profile at regular time intervals. Additionally, the plots include the values in the CV space for the native reference structures. In the case of the path collective variables, values of those structures that were used to build the CV, namely as frames for the CV definition, appear in the plots as golden lines. Similarly, in the plots of CV3 and

CV4 useful reference values are shown respectively in gold and turquoise in the upper panel, and in gold in the lower panel. In all other cases, the native CV value is traced in gray. It can seem surprising to see, in some cases, that the lines corresponding to the native structures fall close, yet not inside the native minimum that we associate to them. This is particularly evident for path-based collective variables and can be explained by recalling that a crystallographic protein structure always expresses an average conformation and, in this particular case, each native structure included in the topology represents an ideal extreme, that results from relaxation and minimization in vacuum. Being a protein inherently dynamic, it is definitely only a fraction of the native contacts, even of those from the same conformation, that is maintained under the influence of a Langevin dynamics at 310 K (our choice to computationally mimic the surrounding biological environment). Since path-based variables depend on a large number of atomic distances deriving from a static conformation, this measure is particularly sensitive to every native contact that detaches from its static definition. In consideration of this, we consider this aspect not problematic and nonetheless consider these structures as highly representative for the underlying topology, and graphically indicative of a tendency.

4.4.2 Comparative analysis of one dimensional free energy profiles

Finally, in the following we expose the results of the comparison of the homologous free energy profiles of all systems. For the sake of clarity, I discuss each structural element separately. It is most convenient to start from the discussion of the auxiliary domains because their arrangement enters the discussion of other variables.

Regulatory domains Simulations of the multidomain constructs reveal an essential trait added by SH3 in the dynamics of c-Abl: a drop in the frequency of SH2 ligation to the N-lobe. A primary role in this issue is played by the strength of intermolecular interactions between SH3 and the linker described in §4.2.2. These interactions lock the strand SH3-SH2-linker in a folded, hairpin-like arrangement and consequently confines the domains to a limited range of movement. It is only when these interactions are released that the linker can extend completely and that the N-lobe interface enters within reach of SH2. When, on the other hand, interactions between linker and SH3 are

effective, SH2 experiences a heavy and indirect restraint to its motion. This limitation acts, if not exclusively by pressing SH2 onto the C-lobe interface, by restricting the motion range around it and creates an overall statistical advantage for the autoinhibited arrangement. In a less idealized picture, molecular crowding should also be counted as an additional factor affecting the domains' movement between the transition extremes. These observations align with the idea of a protein design that is strongly oriented towards autoinhibition by enforcing conditions that efficiently hinder the fully active conformation at a low energetic and evolutionary cost. In this light, then, the comparative scheme we are presenting elucidates a powerful dynamical artifice through which c-Abl achieves the strict downregulation central to its physiology.
Both CV3 and CV4 (in fig. 20 and 21) give a clear description of the dynamics of the system and map into clear free energy profiles with multiple minima that are easily associated to the conformations of interest (the inactive conformation with SH2 at the C-lobe, at low values of CV3 and high values of CV4, the active conformation with SH2 at the N-lobe, vice versa, and finally a large entropic minimum at the largest distance for both CVs that corresponds to all other conformations that hold no functional relevance to the extent of SH2 arrangement). The inequality in sampling of SH2 at the N-lobe we alluded to for the previous collective variables emerges most clearly here.
From a quantitative point of view, the balancing between the minima depends critically on the exhaustive sampling of the free energy space and drawing quantitative conclusions seems premature at this point.

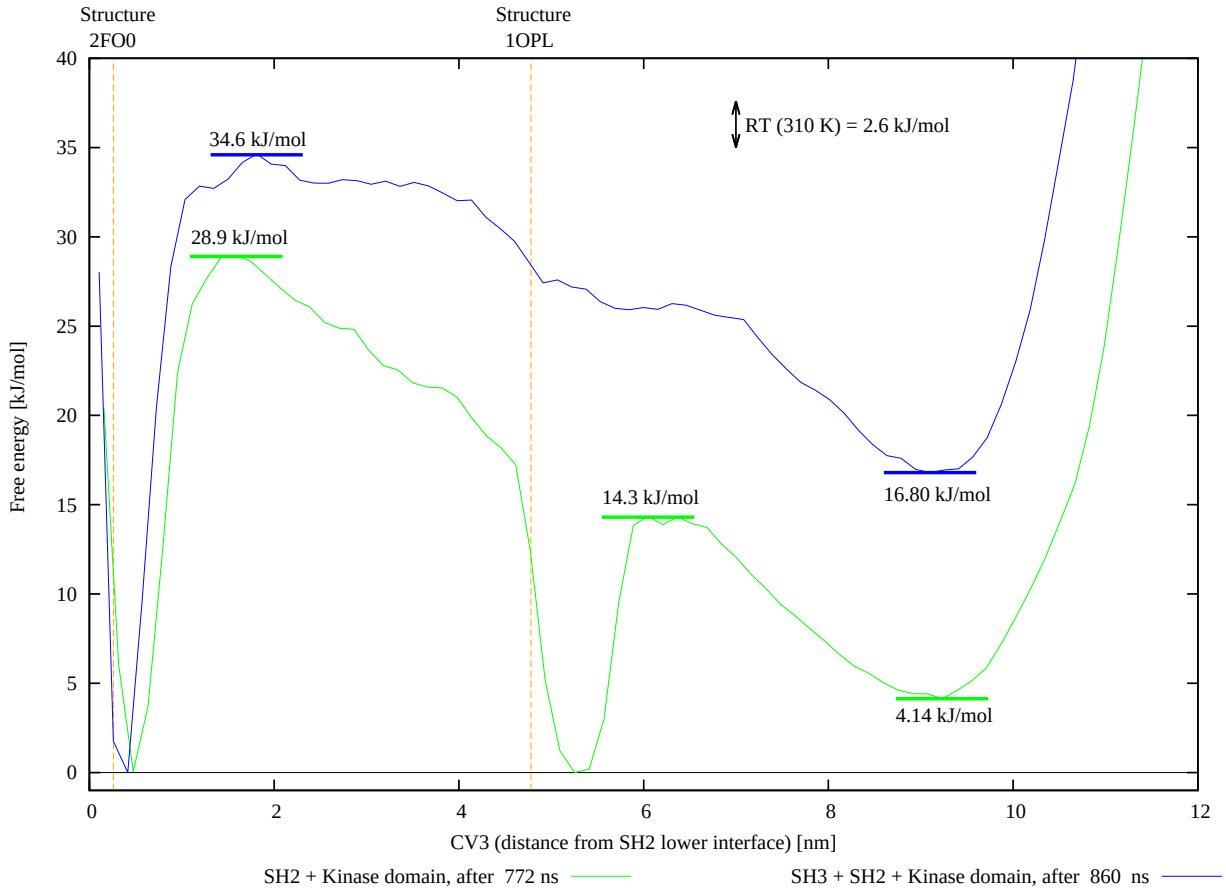

Figure 20: Comparison of the final free energy profiles for CV3 (distance from SH2 lower interface).

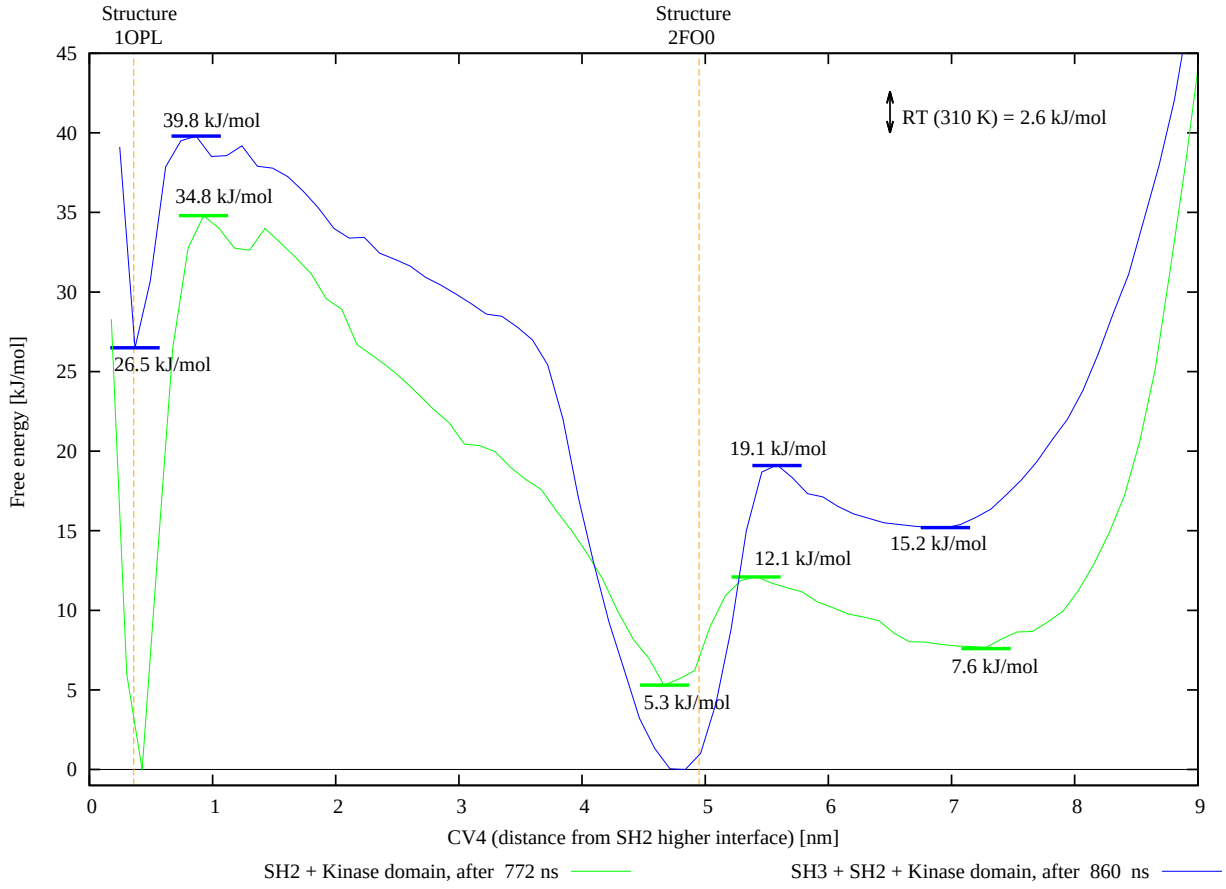

Figure 21: Comparison of the final free energy profiles for CV4 (distance from SH2 higher interface).

ACTIVATION LOOP Comparison of the three free energy profiles for the activation loop (fig.22 and table 7) shows that, while in the kinase domain the open conformation has a higher free energy of 6.3 kJ/mol than the closed one, the difference in free energy between the two basins $\Delta\Delta G_{\text{open}\to\text{closed}}$ falls for both around zero kJ/mol when at least one auxiliary domain is included. The difference between $\Delta\Delta G_{\text{open}\to\text{closed}}$ for the two systems is within RT (2.6 kJ/mol) (here and in the following we take RT as the minimal threshold to consider reliable a difference in free energy). Keeping in mind the caution of the RT threshold, we also remark, when only SH2 is present, a slight inversion in the balance between conformations. This may hint to a change of global minimum, with the open loop taking this role, at variance with other systems in which the closed one seems to be statistically preferential.
The energy barrier for the converged calculation of the kinase domain is found at 8.9 kJ/mol. For the multidomain systems, these values are 8.6 kJ/mol for the second system (SH2+kd) and 13.5 kJ/mol for the largest one.

	kd	SH2 + kd	SH3 + SH2 + kd
$\Delta\Delta G_{\text{open}\to\text{closed}}$ [kJ/mol]	-6.3	0.7	-0.8
Barrier height from global minimum [kJ/mol]	8.9	8.6	13.5

Table 7: Highlights of the comparison of the final free energy profiles for CV1 (s-path on aloop), shown in fig. 22.

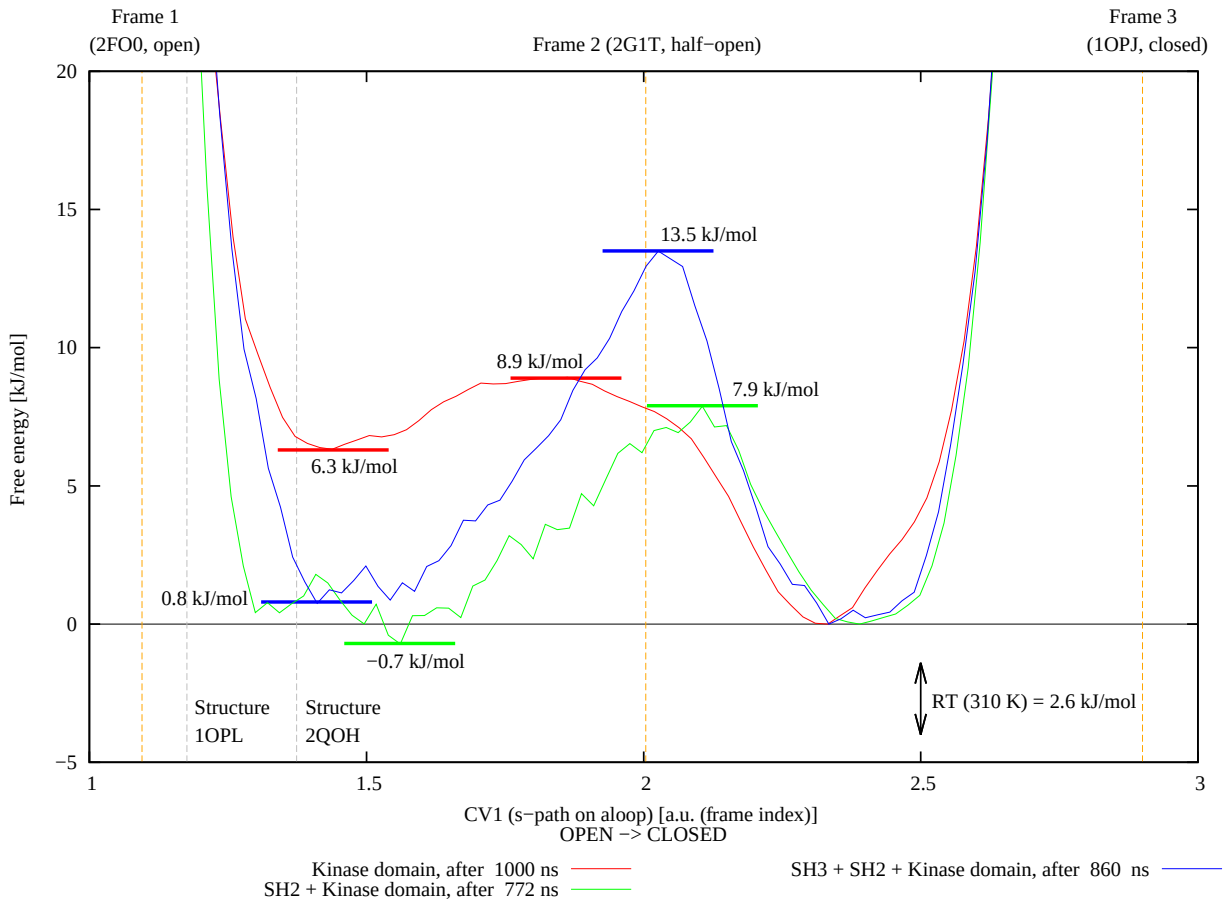

Figure 22: Comparison of the final free energy profiles for CV1 (s-path on aloop).

DFG MOTIF The DFG-in state steadily occupies the global minimum in all simulations and is then consistently recovered as the most favorable conformation with respect to the DFG-out. This is clearer in the free energy profile of the first extracted and reweighted variable (the distance Ile 312 CG1 - Asp 400 CG, fig. 23). The path variable CV2 rather served successfully to induce the DFG transition (fig. 24) in the metaD framework. This finding agrees qualitatively with all previous indications from classical force field and other non-structure-based methods (see page 41) on the balance between the two DFG orientations. Again, engagement of the SH2 and SH3 has a remarkable effect on the populations that results in free energy differences beyond RT. Only SH2 + kd exhibits a minimum for the DFG out conformation (with $\Delta\Delta G_{\text{in}\rightarrow\text{out}} = 4$ kJ/mol), where the kinase domain reveals a plateau at 9.9 kJ/mol and the system starting at SH3 shows a descending and rather steep profile at 13.0 kJ/mol at the same coordinate. Here, then, we face divergent outcomes, leaving the profile for the kinase domain lying halfway between the other two. Given the design of this experiment, this result can easily reflect the different engagement of the adaptor domains, although this remains to be confirmed at the conclusion of the calculations. It is clear in fact that SH2 + kd spans the transition state for the SH2 displacement more effectively than the third system and this fact may be having an impact here. In this sense, our data support an higher occurrence of the DFG flip and a stabilization of the DFG Asp out orientation in the case of a more efficient drift of the regulatory domains. This notion will find further support when we will discuss the motion of the lobes' hinge.

	kd	SH2 + kd	SH3 + SH2 + kd
Energy at DFG Asp *out* [kJ/mol]	plateau 9.9	$\Delta\Delta G_{\text{in}\rightarrow\text{out}}$ -4.0	descending curve 13.0
Barrier height from global minimum [kJ/mol]	/	9.6	/

Table 8: Highlights of the comparison of the final free energy profiles for the DFG, shown in fig. 23 and 24.

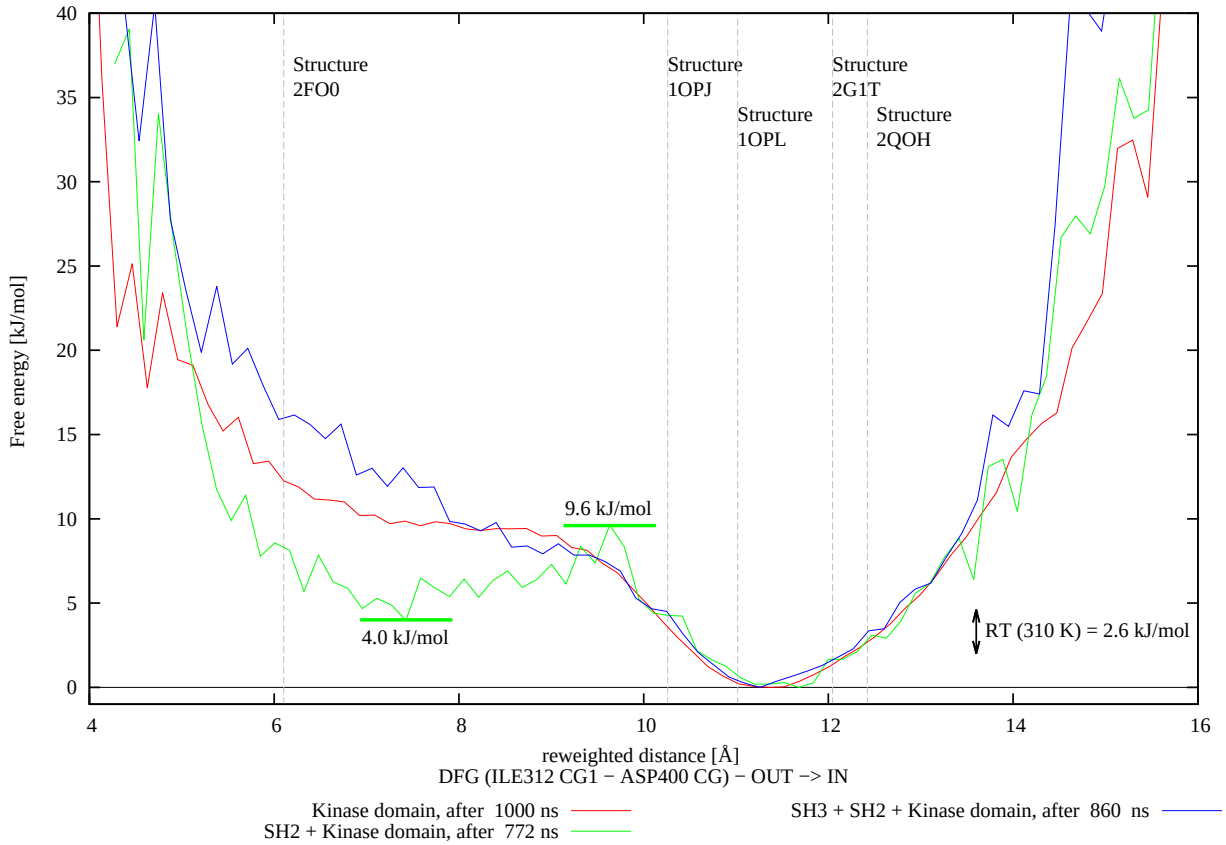

Figure 23: Comparison of the final free energy profiles for the extracted distance ILE312 CG1 - ASP400 CG (DFG, out $\rightarrow$ in).

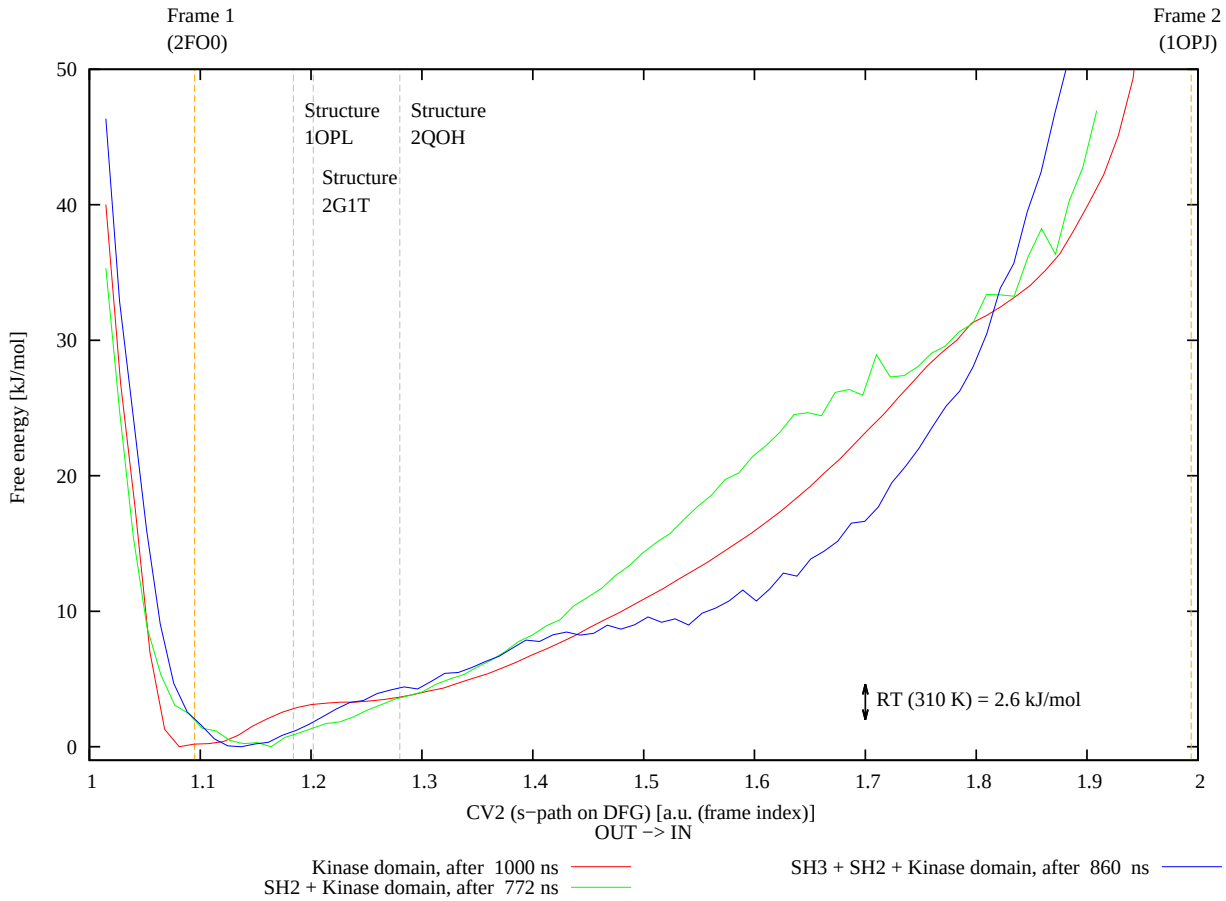

Figure 24: Comparison of the final free energy profiles for of CV2 (s-path on DFG).

ABL-LIKE AND SRC-LIKE INACTIVATION MODES Two distances were extracted and reweighted to investigate the concerted path that critical amino acids follow when assuming the active and the inactive conformations referred to as Abl- and Src-like during its catalytic cycle, and to reveal if this conformational transition is functionally coupled to others in the domain arrangement.
Regarding the fundamental ion pair between Lys 290 in the active site and Glu 305 from the αC helix (fig. 25), neither a clearly defined transition state nor any meaningful difference among the profiles can be distinguished and we conclude that this observable appears unaffected by the other structural factors we considered.
With respect to the salt bridge with Arg 405 that exposes Glu 305 to the solvent in the Src-like inactive conformation (figure 26), the SH2+kinase domain system profile leans towards the extreme of the transition at low distance, resulting in a significantly lower extreme at short distance with respect to the kinase domain and SH3+SH2+kd. This points to an influence of SH2 in this motion, in the sense that SH2 at the N-lobe or unligated may disfavor the salt bridge in question.

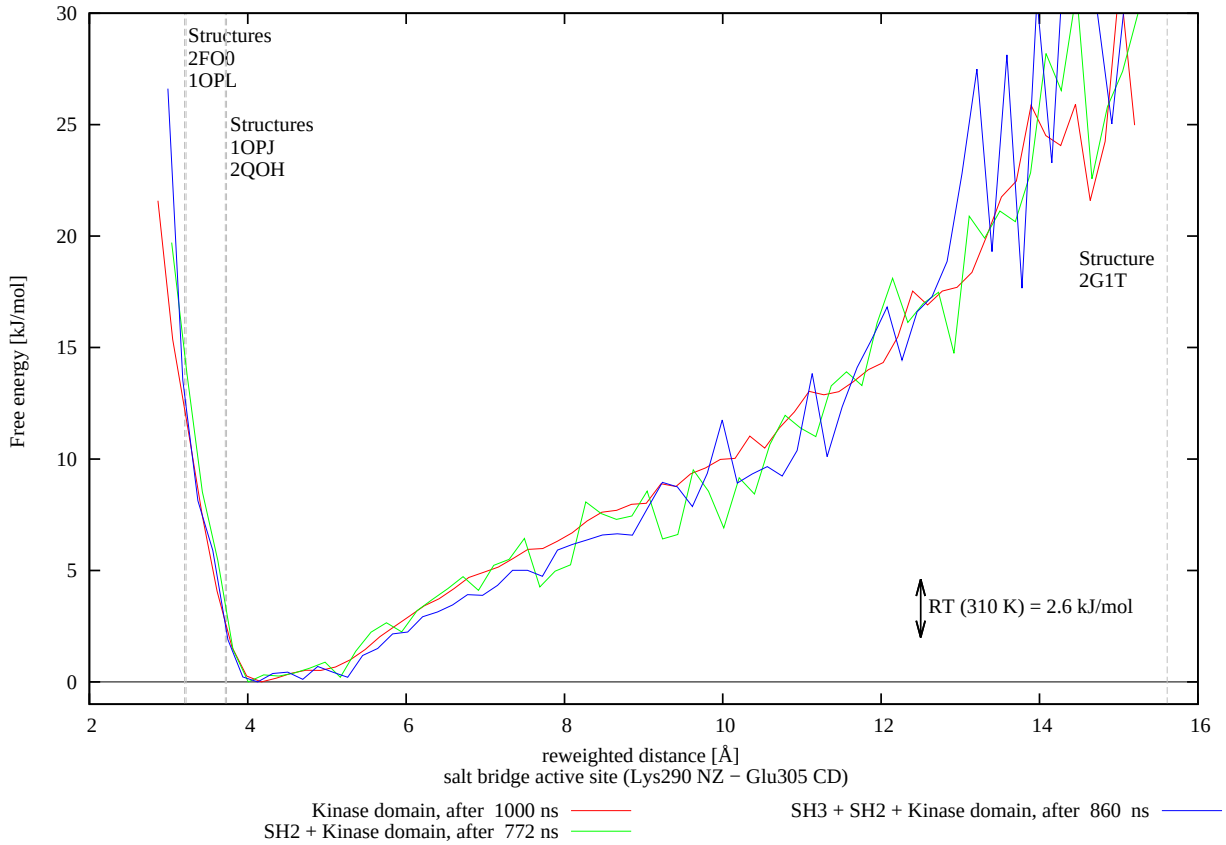

Figure 25: Comparison of the final free energy profiles for extracted distance Lys 290 NZ - Glu 305 CD (active site, active → Abl-like inactive).

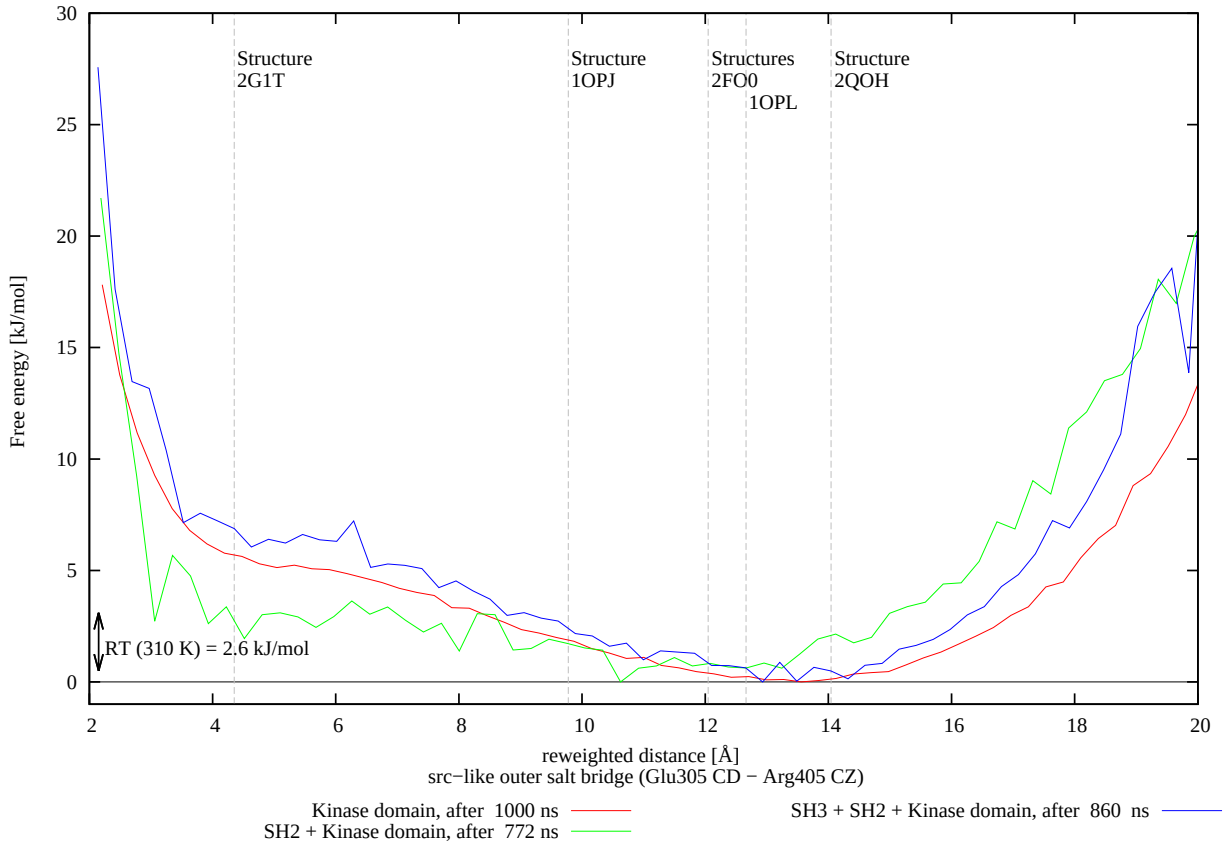

Figure 26: Comparison of the final free energy profiles for extracted distance Glu 305 CD - Arg 405 CZ (external salt bridge, Src-like inactive $\rightarrow$ any other state)

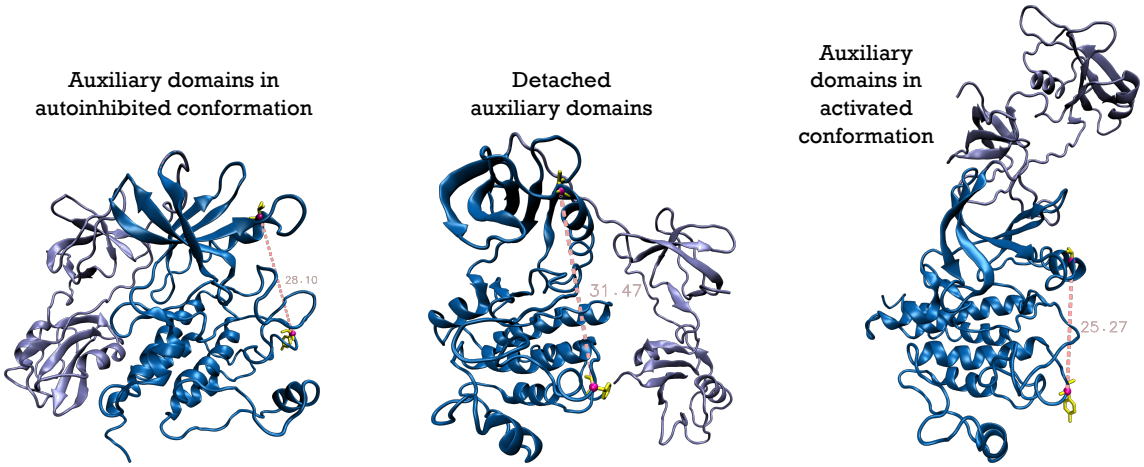

Figure 27: Examples of lobes separation, a quantity rooted in the hinge flexion that was measured through the distance Val 299 CA - Tyr 432 CA. These atoms are highlighted in fuchsia and the relative residue is shown in yellow. A possible correlation with the position of the auxiliary domains emerges in the discussed simulations.

HINGE MOTION To measure this choral and subtle movement described at page 41, the distance Val 299 CA - Tyr 432 CA was chosen (see fig. 27 for a visualization and fig. 28 for the final free energy plots). Even at eyesight is evident that the average distance spanned by SH3+SH2+kd is lower than for the other systems. In fact, the width of its profile when crossing an ideal reference line of 5 kJ/mol is of 5.7 Å against 7.3 Å for both other systems. Thus, again in agreement with the experimental indications, the hinge motion is indeed limited when the autoinhibited conformation is most often visited. Moreover, the hinge motion curve leans towards lower values for SH2+kd in comparison with other systems. The presence and behavior of SH2 may then contribute to reduce the average separation between the lobes. As a result, another contribution to c-Abl autoinhibitory strategy has been consistently revealed by the model and shows potential of further insight into c-Abl machinery.

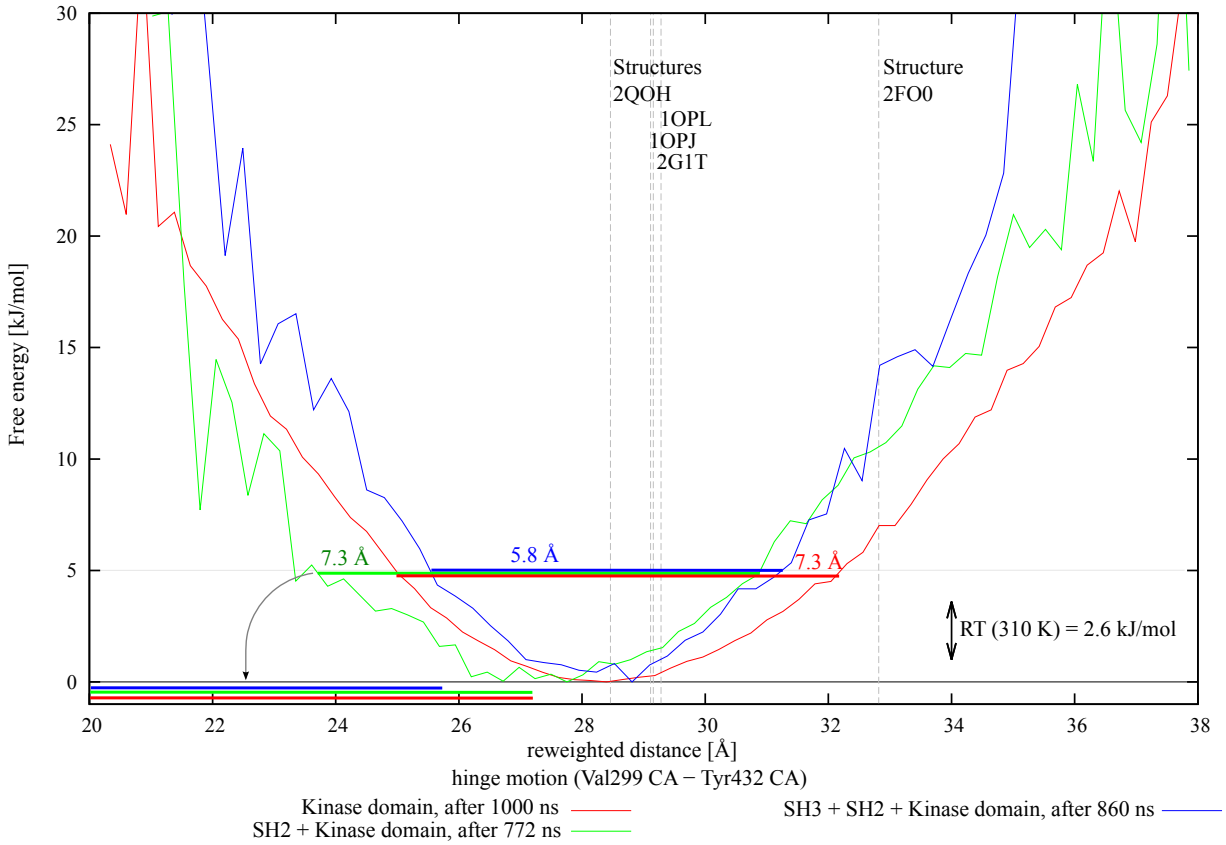

Figure 28: Comparison of the final free energy profiles for extracted hinge motion measure (distance Val 299 CA - Tyr 432 CA).

4.5 CONCLUDING REMARKS

The computational approach and the application we described compose known experimental notions and theoretical concepts with the goal of describing the interplay of conformational transitions driving c-Abl. To the extent allowed by the advancement of the simulations, we have shown that the presence, localization and dynamics of the auxiliary modules results in quantitative and qualitative differences in the free energy profiles corresponding to conformational changes in several structural elements in c-Abl. Their contribution has primary importance in determining the key balance of populations needed for proper function, and represents a proof of concept of the crucial, quantifiable relevance of the regulatory domains in the action of the catalytic core they serve.
Here we summarize the conclusions reached in this discussion:

- The articulate movement of the auxiliary domains in this simulations elucidates in dynamical details the primary role of the SH3 domain in c-Abl autoinhibition: by establishing intermolecular interactions with the linker, it keeps the whole auxiliary apparatus (composed by the linker + SH2 + SH3) blocked in a limited range of movement around the autoinhibiting conformation.

- For the system featuring SH2 but not SH3, with the previous effect thus removed by construction, an highly enhanced detachment and drift of the auxiliary gear (SH2 and linker) was revealed, that cascades in enhanced binding to the activating N-lobe interface and distinctly evidences the downregulating effect of SH3 in full length c-Abl.

- When SH2 is bound to the C-lobe, the lobes average separation is consistently low. As a consequence, the results also let infer that SH2 detachment from the C-lobe interface appears associated to other transitions: the activation loop opening and closing, the DFG flip and hinge motion opening.

- No specific correlation was isolated in the salt bridges underpinning Src and Abl like inactivation.

The combination of the structure-based approach and parallel tempering metadynamics sampling boosted the insight potential of the model to its maximum.

By rationally choosing and merging available informations, we built a tool whose descriptive power goes beyond the sum of its elements and gained unique insight into protein regulation. In perspective, this approach holds the promise of helping the rational description and the revelation of still hidden vulnerabilities in the regulation of whole proteins, and to contribute in this way to advance biomedical progress.

CONCLUSIONES

El protocolo que hemos descrito anteriormente se compone tanto de nociones experimentales como de conceptos teóricos que se engloban en un esquema innovador, cuyo principal objetivo es describir la interacción de las transiciones conformacionales producidas en la proteína c-Abl. En la medida que el avance de las simulaciones nos permite, hemos demostrado que la presencia, localización y dinámica de los módulos auxiliares añaden diferencias cuantitativas y cualitativas en los perfiles de energía libre con respecto a numerosos elementos estructurales de los sistemas simulados de la proteína c-Abl.

Así mismo se demuestra como la contribución de dichos módulos tiene gran importancia a la hora de determinar el equilibrio de las poblaciones conformacionales que es indispensable para que una proteína exprese correctamente su papel biológico. Y por último, este trabajo representa una prueba de concepto de la relevancia crucial y cuantificable de los dominios auxiliares en la acción del dominio catalítico al que sirven.

Al haber escogido y conjugado racionalmente información públicamente accesible, modelización basada en la estructura y los cálculos de metadinámica a temperaturas paralelas (conocida como *parallel tempering metadynamics*) hemos construido una herramienta cuyo poder descriptivo va más allá de la suma de sus elementos, adquiriendo una visión única de la regulación de la proteína c-Abl.

En perspectiva, este enfoque nos brinda la posibilidad de describir racionalmente la dinámica de proteínas completas, con el fin de revelar vulnerabilidades estructurales en las que se basen tanto las enfermedades como los tratamientos farmacológicos (como en el caso de c-Abl), y ofreciéndonos de esta manera un gran potencial de contribuir a promover el progreso biomédico hacía nuevos éxitos.

APPENDIX

EVOLUTION OF THE TRAJECTORIES AT 310 K

The following plots refer to §4.4.1, where an overview on the contents of this appendix can be found. These plots show, in the upper panel, the monitoring of each CV of the three trajectories, and the evolving free energy surface that results from it in the lower panel.

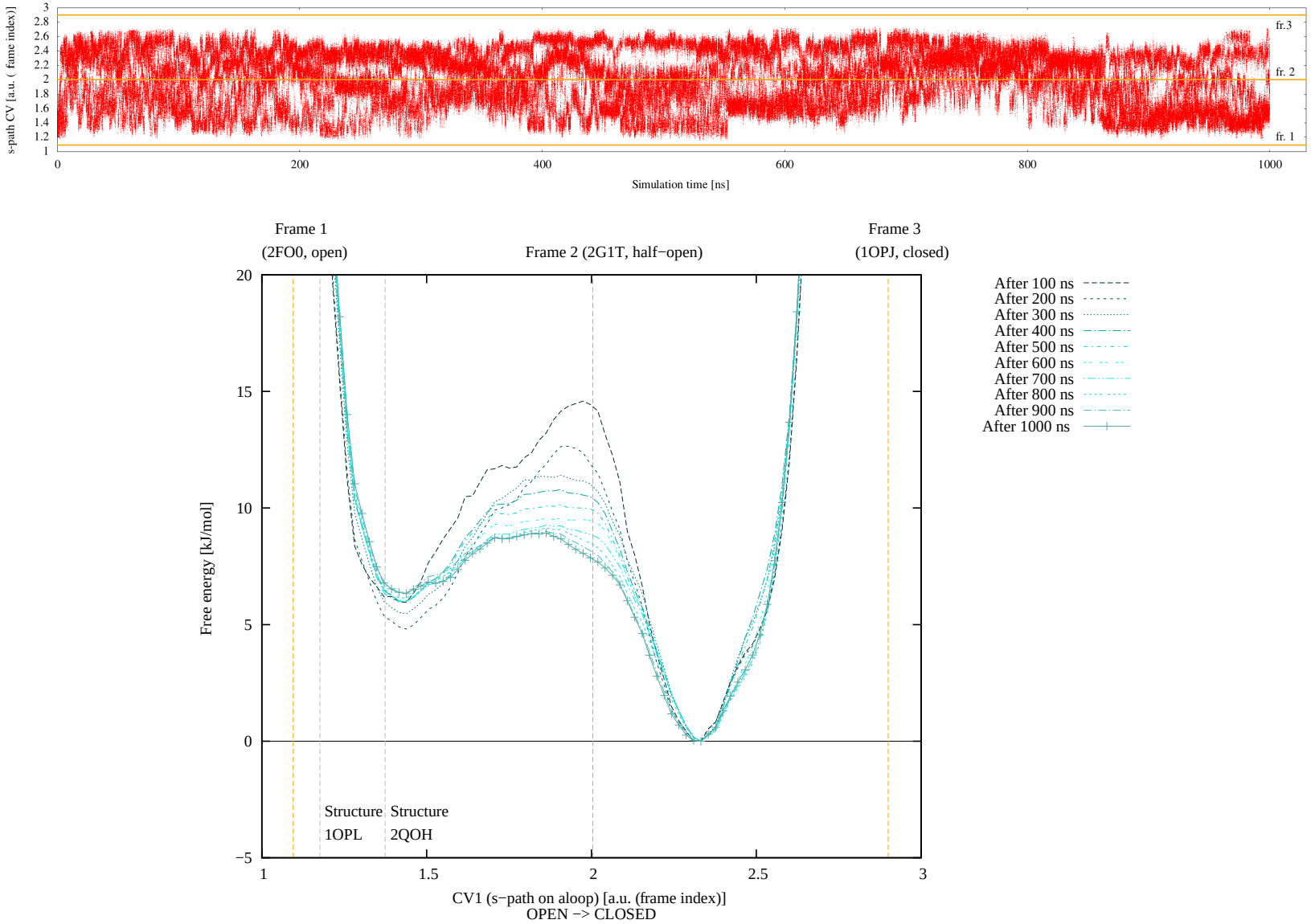

Figure 29: Kinase domain. Evolution along simulation time of CV1 (s-path on aloop, upper panel) and of its free energy landscape (lower panel) at 310 K.

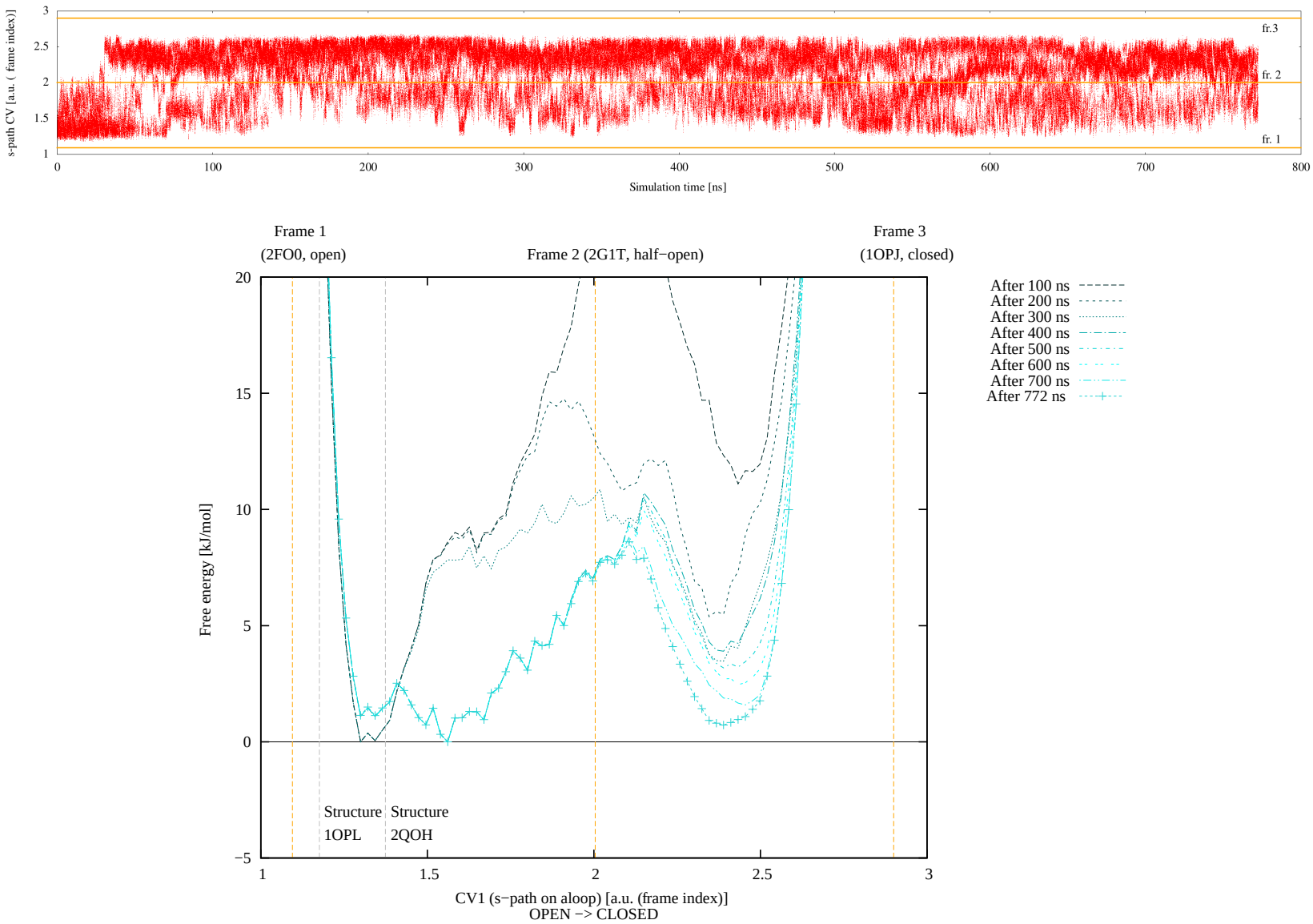

Figure 30: SH2 + Kinase domain. Evolution along simulation time of CV1 (s-path on aloop, upper panel) and of its free energy landscape (lower panel) at 310 K.

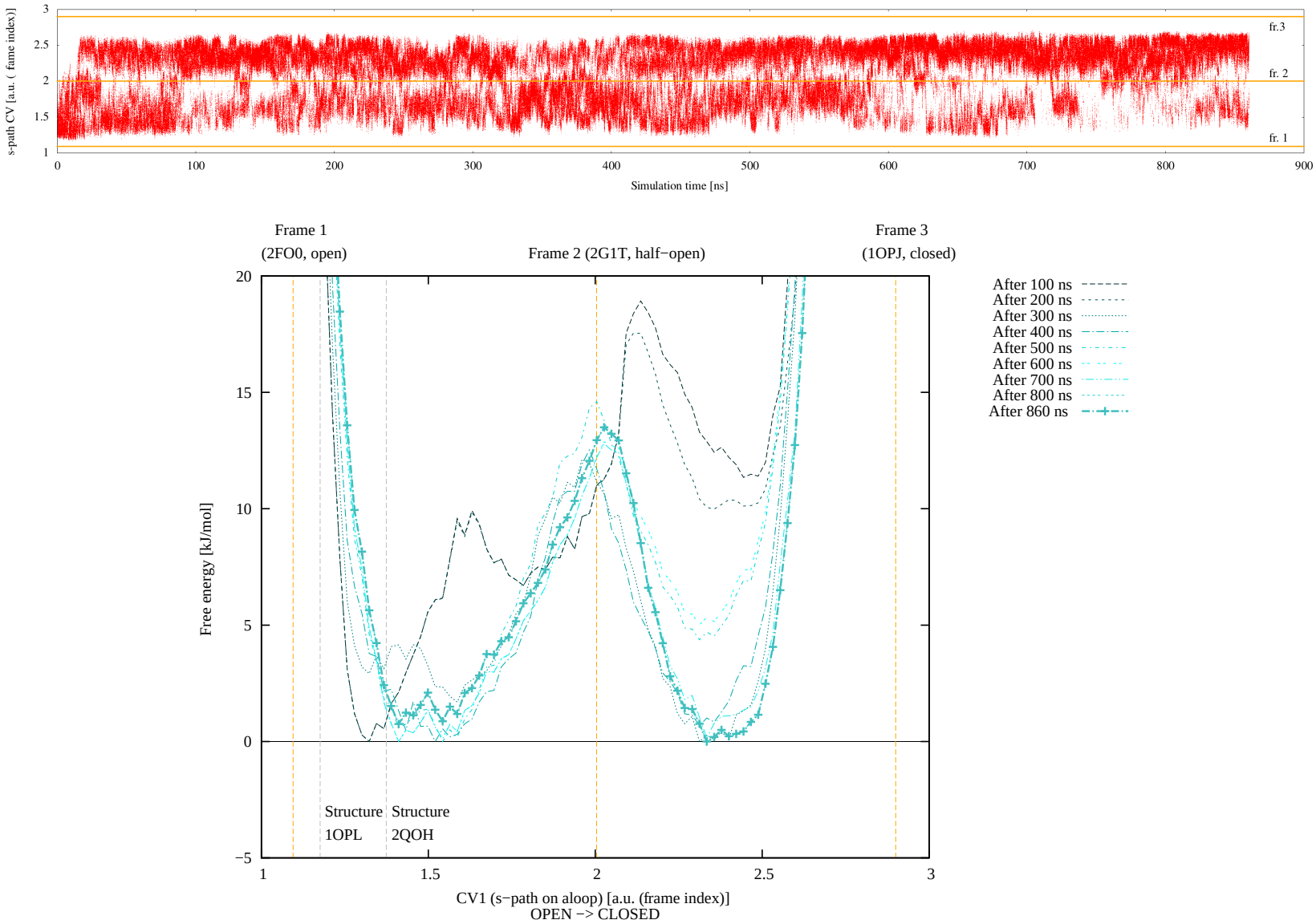

Figure 31: SH3 + SH2 + Kinase domain. Evolution along simulation time of CV1 (s-path on aloop, upper panel) and of its free energy landscape (lower panel) at 310 K.

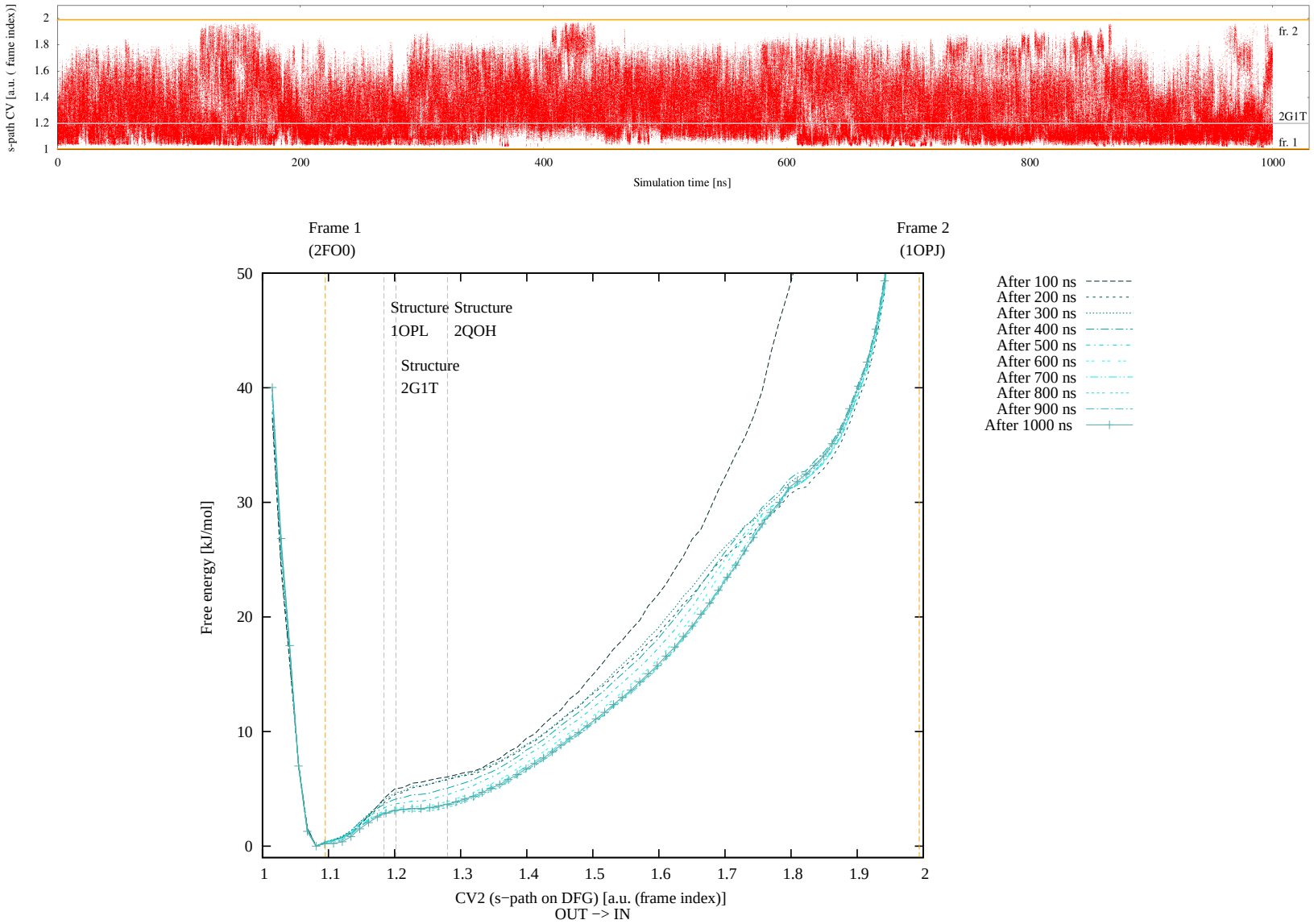

Figure 32: Kinase domain. Evolution along simulation time of CV2 (s-path on DFG, upper panel) and of its free energy landscape (lower panel) at 310 K.

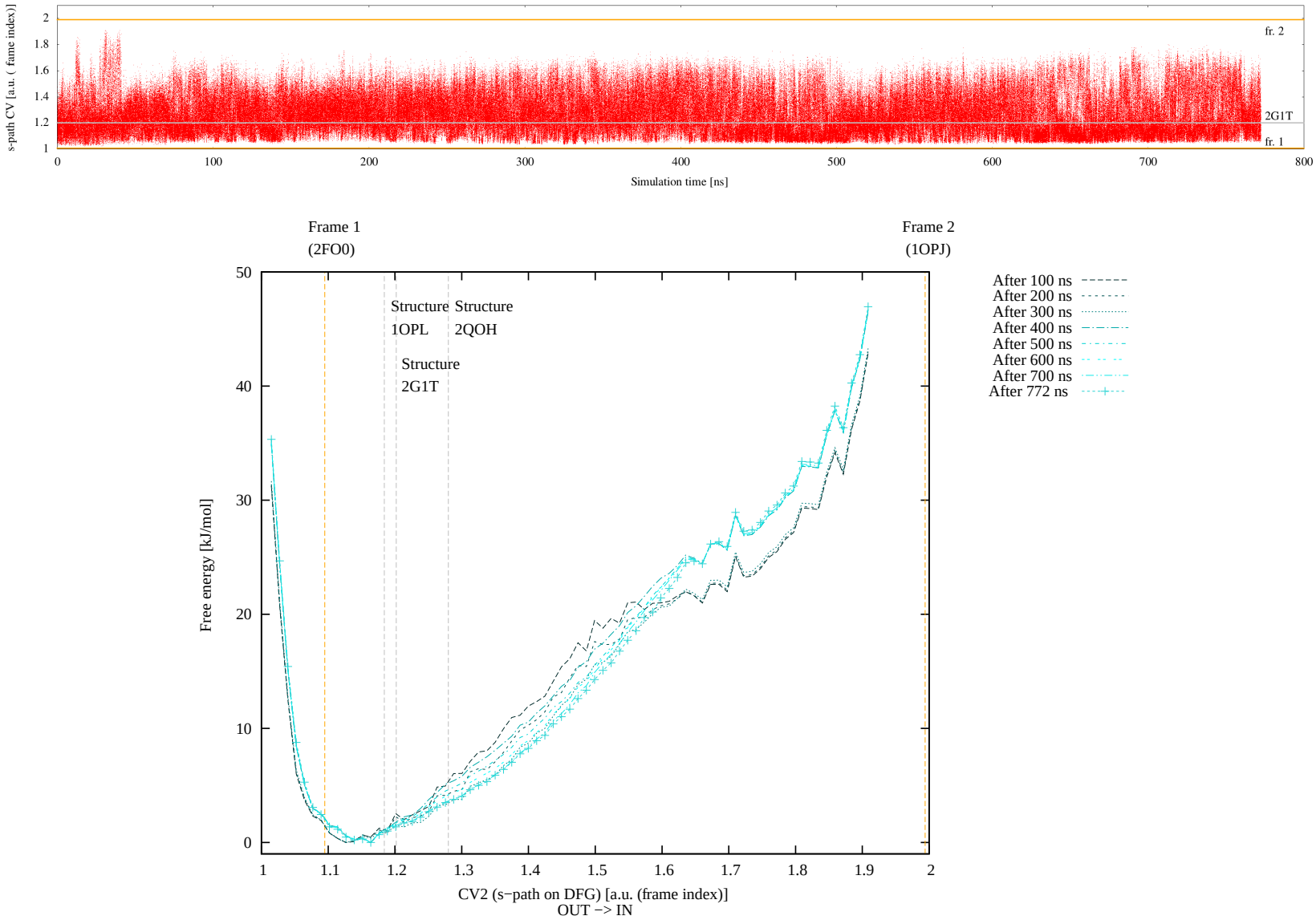

Figure 33: SH2 + Kinase domain. Evolution along simulation time of CV2 (s-path on DFG, upper panel) and of its free energy landscape (lower panel) at 310 K.

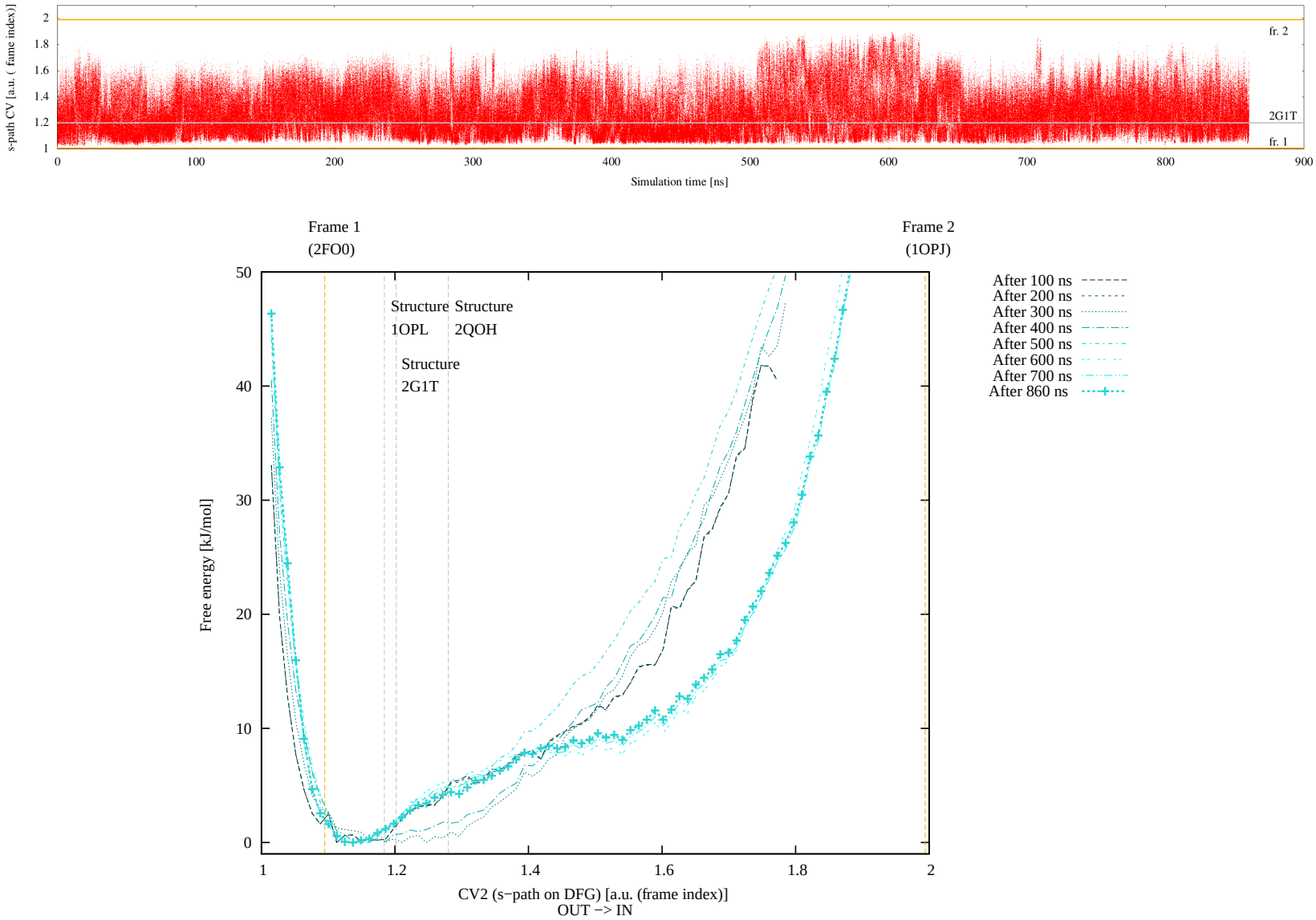

Figure 34: SH3 + SH2 + Kinase domain. Evolution along simulation time of CV2 (s-path on DFG, upper panel) and of its free energy landscape (lower panel) at 310 K.

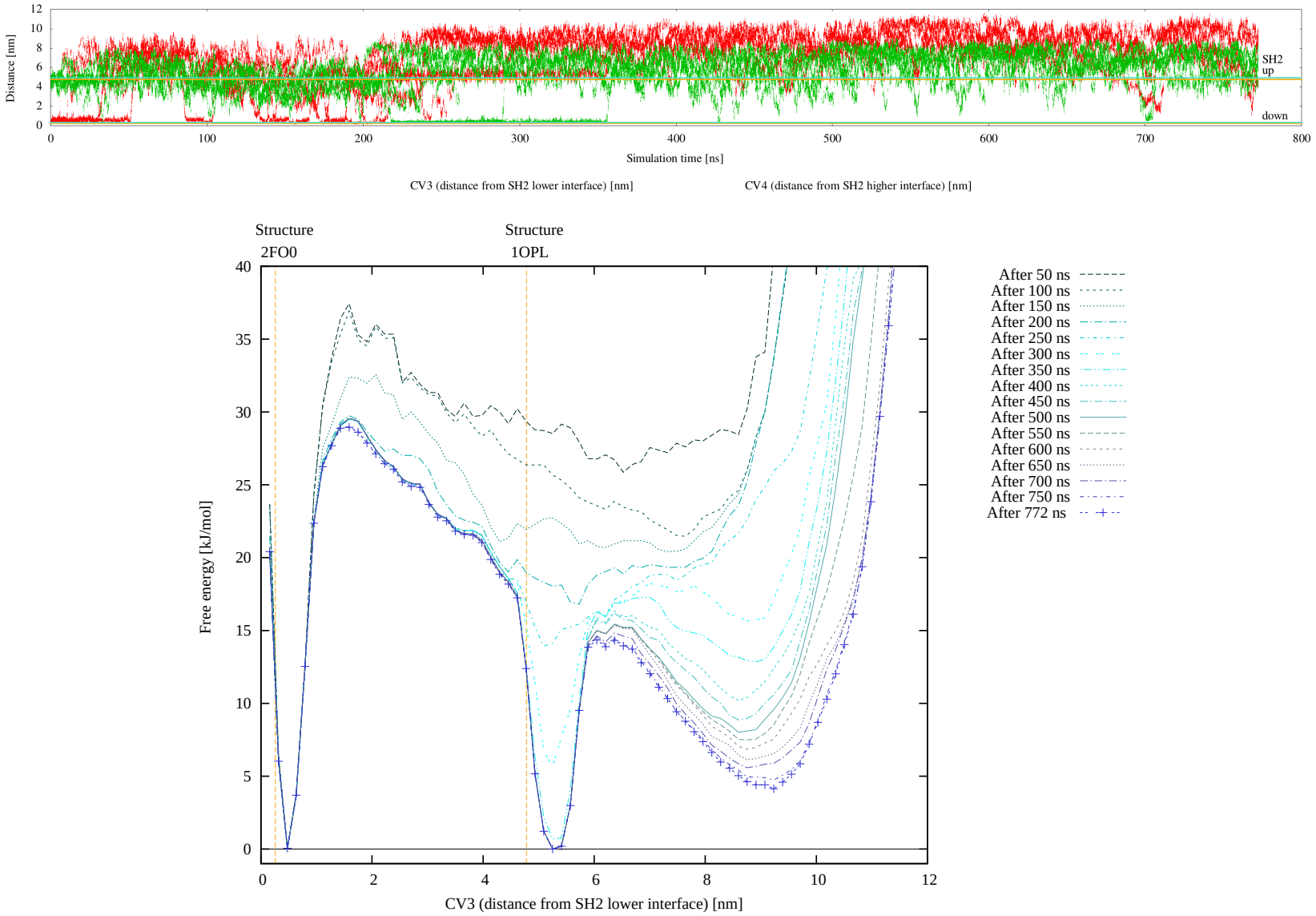

Figure 35: SH2 + Kinase domain. Evolution along simulation time of CV3 (distance from SH2 lower interface, red in the upper panel, where CV4 is also shown in green) and of its free energy landscape (lower panel) at 310 K.

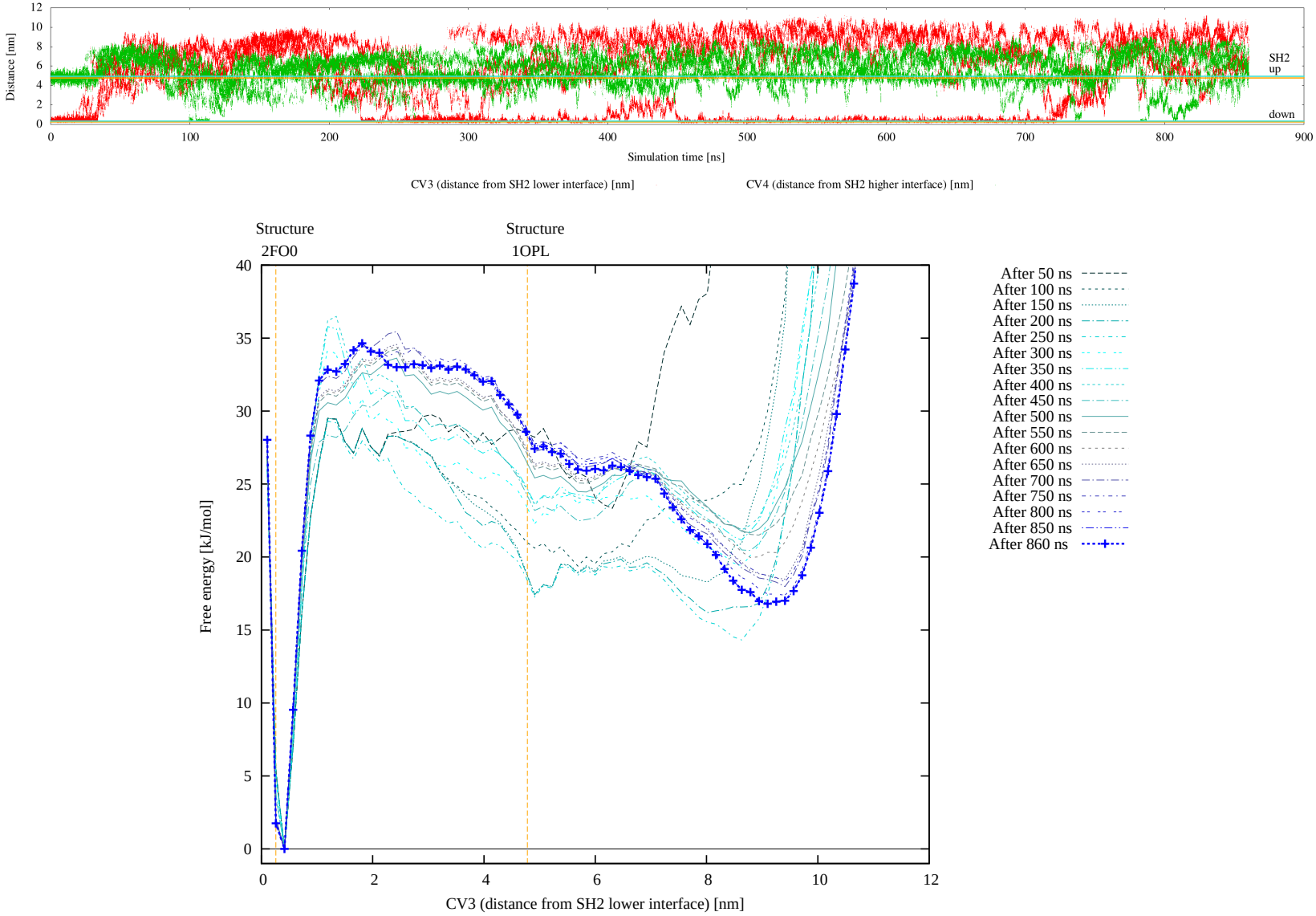

Figure 36: SH3 + SH2 + Kinase domain. Evolution along simulation time of CV3 (distance from SH2 lower interface, red in the upper panel, where CV4 is also shown in green) and of its free energy landscape (lower panel) at 310 K.

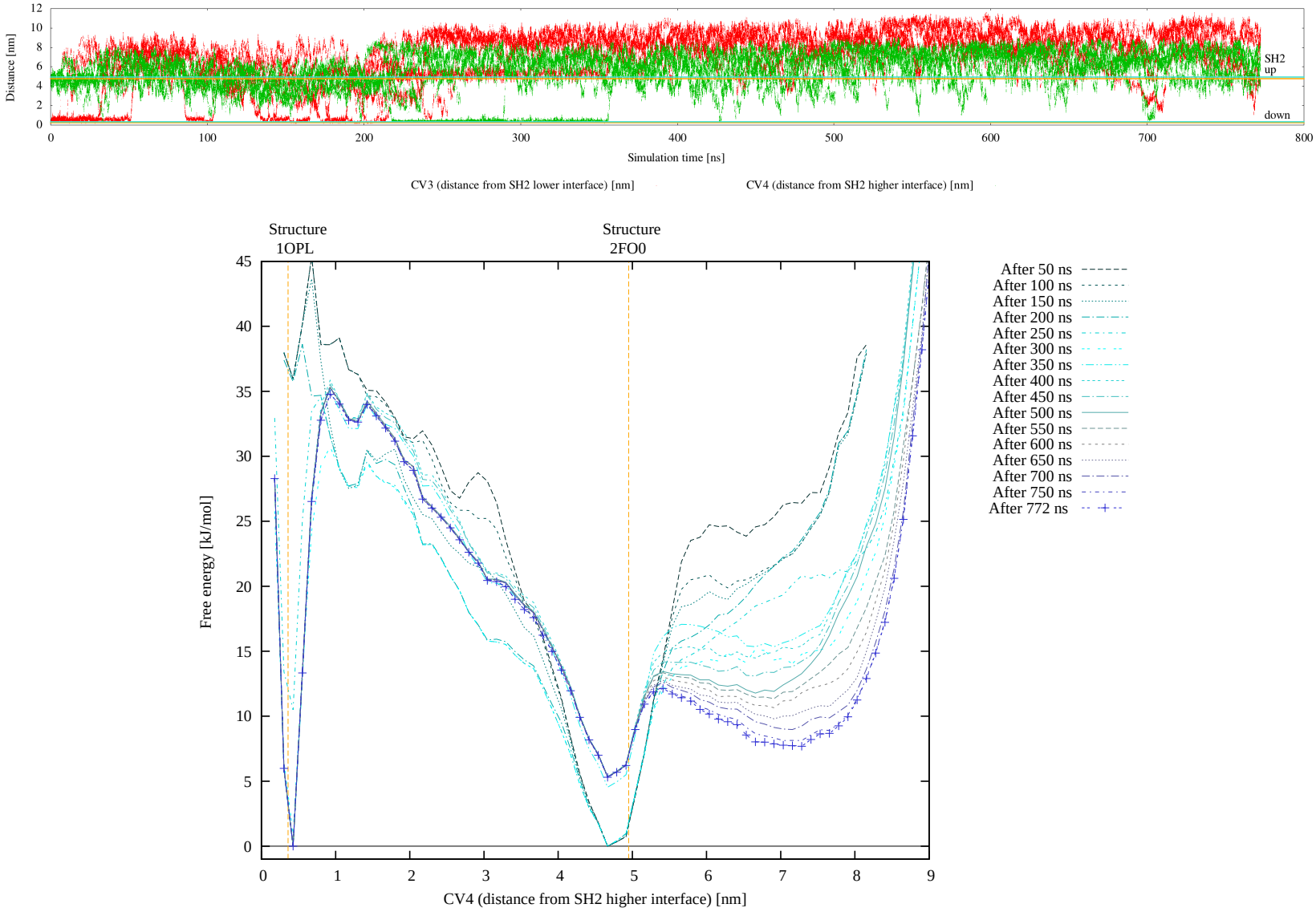

Figure 37: SH2 + Kinase domain. Evolution along simulation time of CV4 (distance from SH2 higher interface, green in the upper panel, where CV3 is also shown in red) and of its free energy landscape (lower panel) at 310 K.

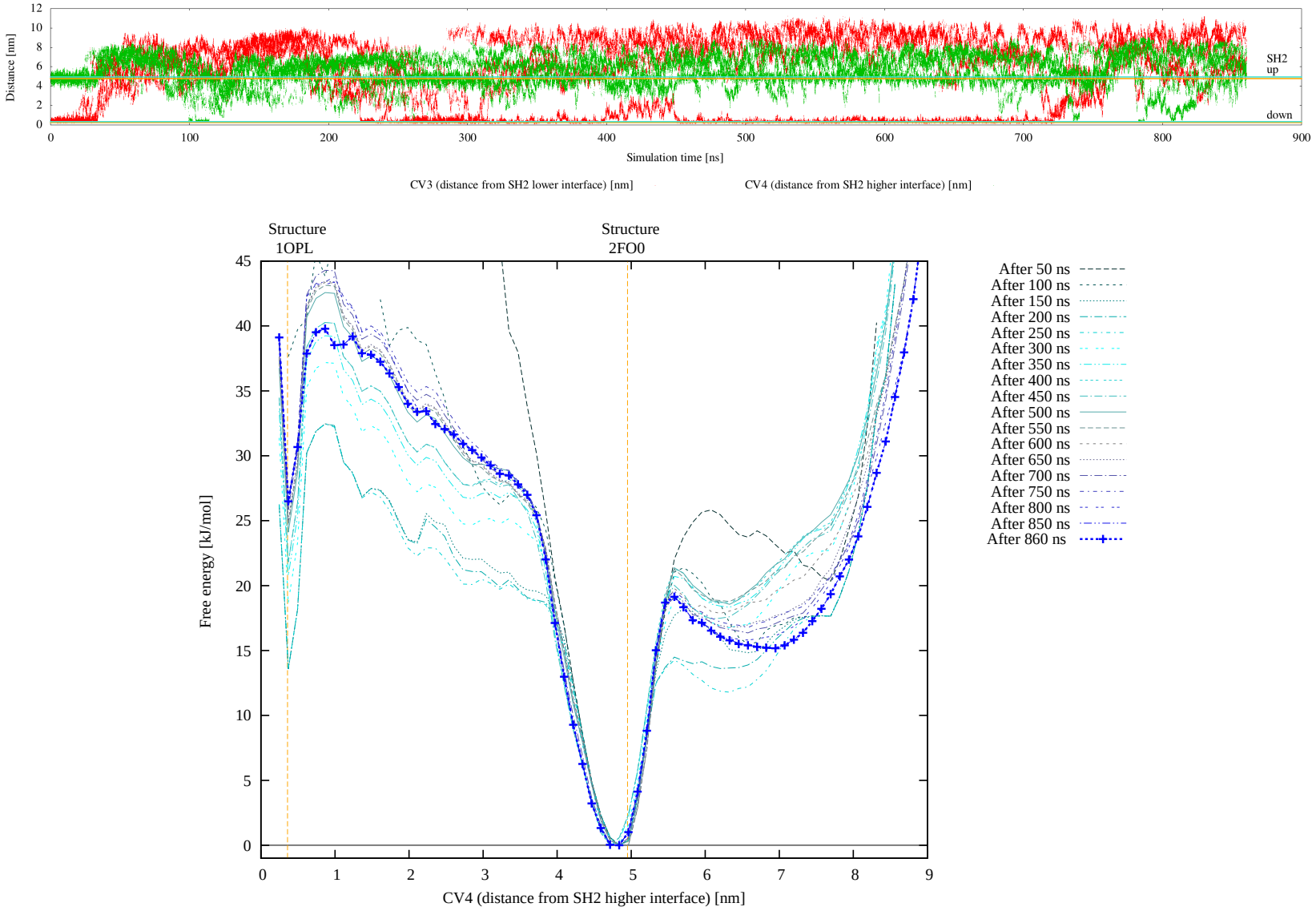

Figure 38: SH3 + SH2 + Kinase domain. Evolution along simulation time of CV4 (distance from SH2 higher interface, green in the upper panel, where CV3 is also shown in red) and of its free energy landscape (lower panel) at 310 K.

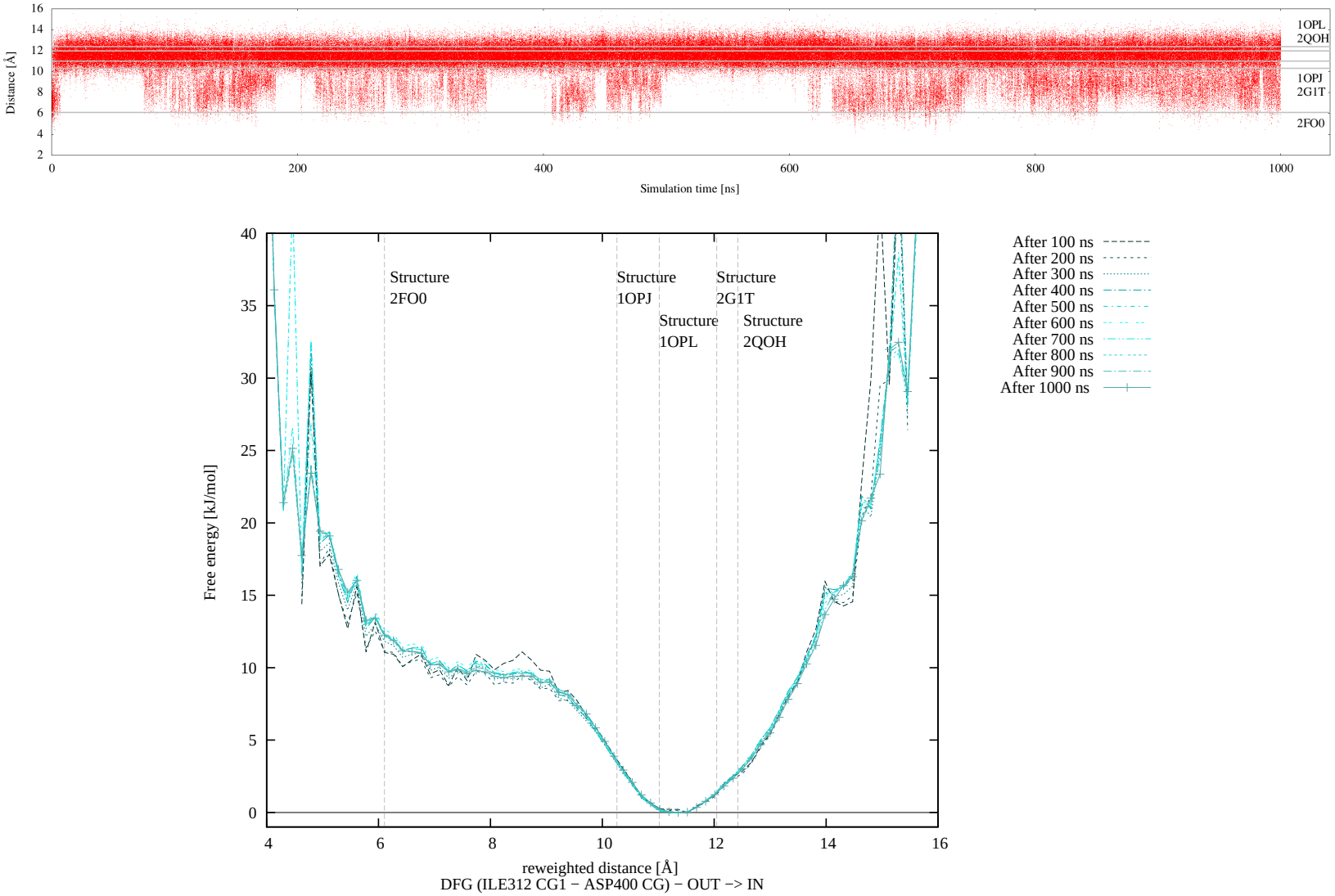

Figure 39: Kinase domain. Evolution along simulation time of the extracted distance ILE312 CG1 - ASP400 CG (DFG, out → in; upper panel) and of its free energy landscape (lower panel) at 310 K.

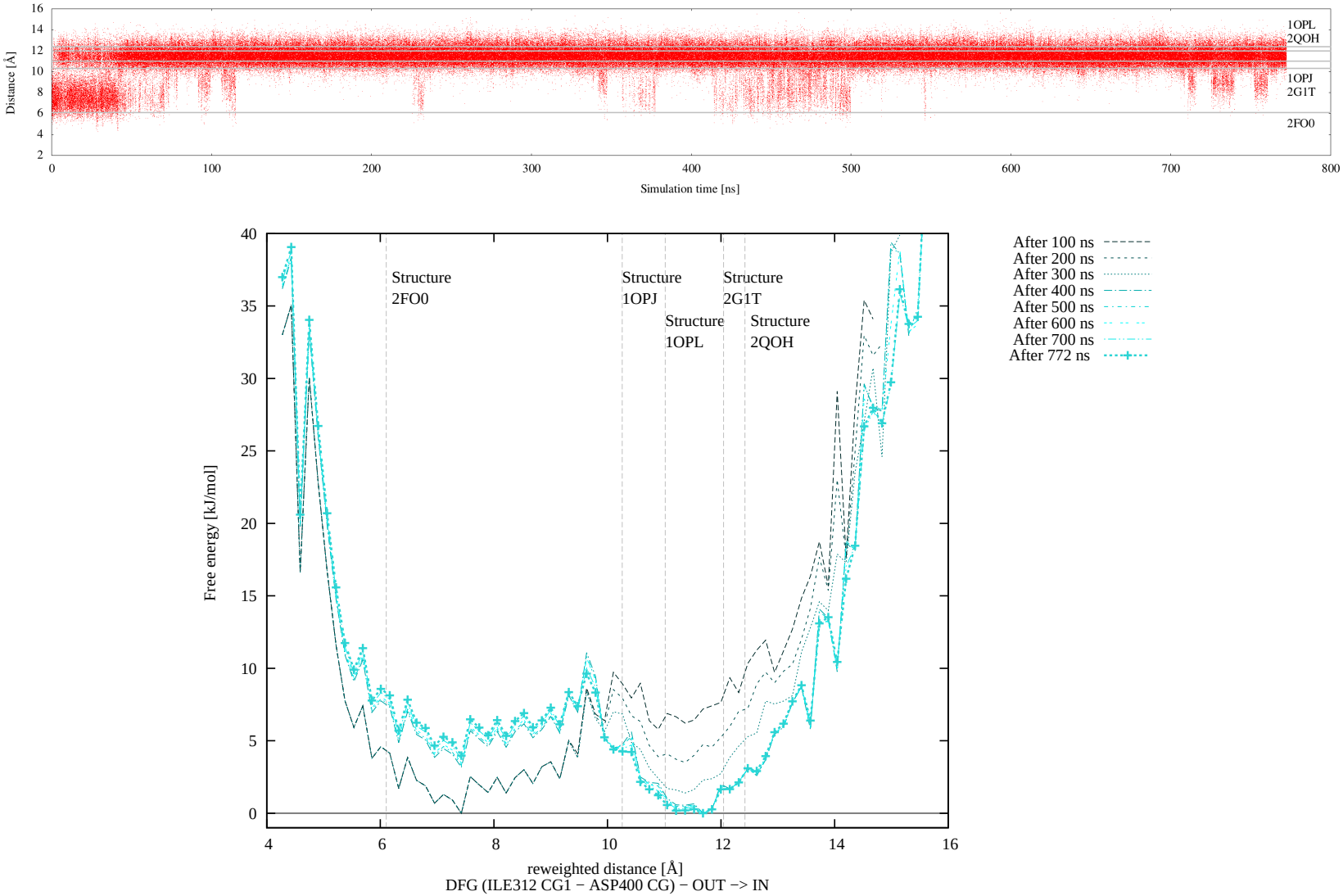

Figure 40: SH2 + Kinase domain. Evolution along simulation time of the extracted distance ILE312 CG1 - ASP400 CG (DFG, out → in; upper panel) and of its free energy landscape (lower panel) at 310 K.

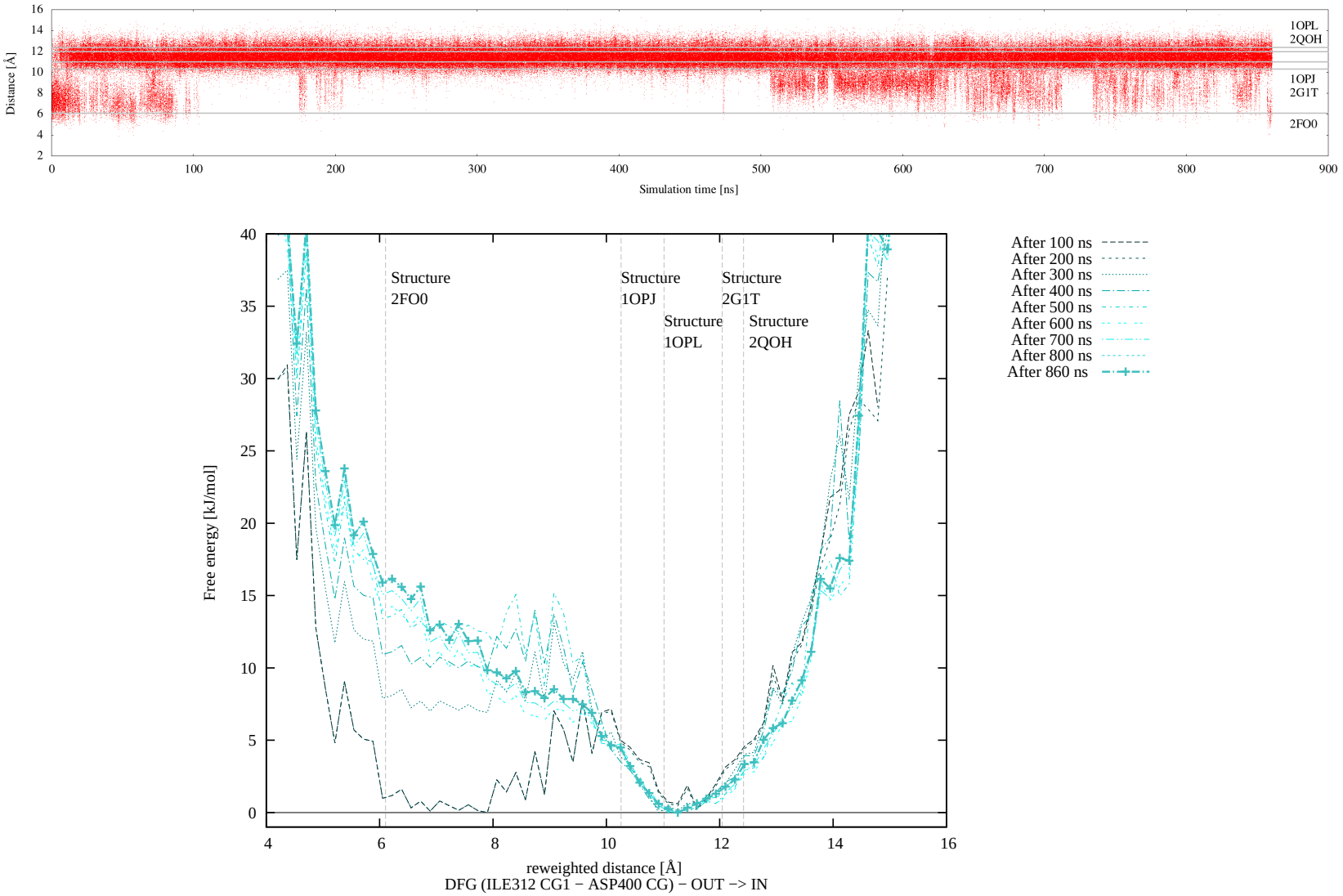

Figure 41: SH3 + SH2 + Kinase domain. Evolution along simulation time of the extracted distance ILE312 CG1 - ASP400 CG (DFG, out → in; upper panel) and of its free energy landscape (lower panel) at 310 K.

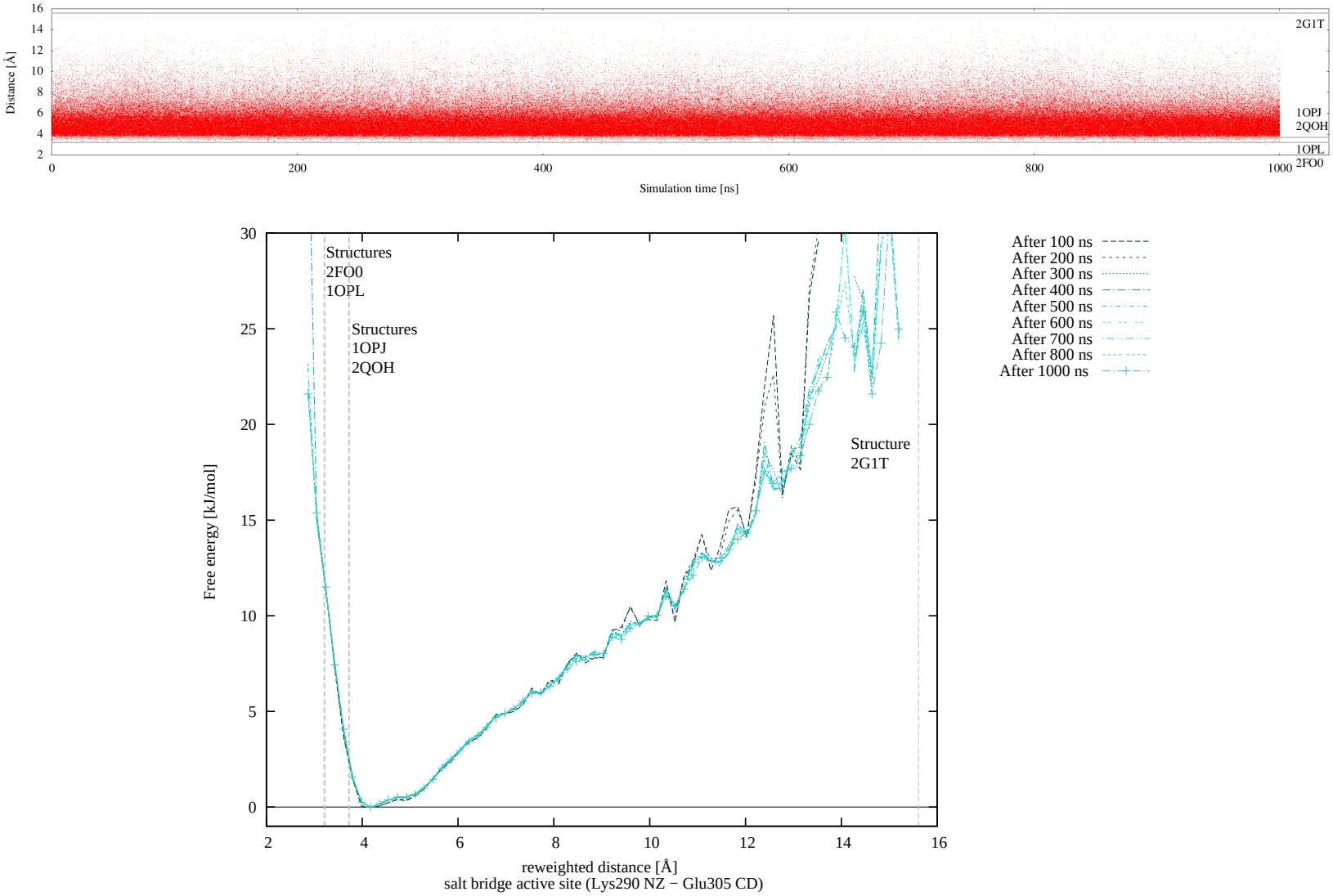

Figure 42: Kinase domain. Evolution along simulation time of the extracted distance Lys 290 NZ - Glu 305 CD (active site, active → Abl-like inactive; upper panel) and of its free energy landscape (lower panel) at 310 K.

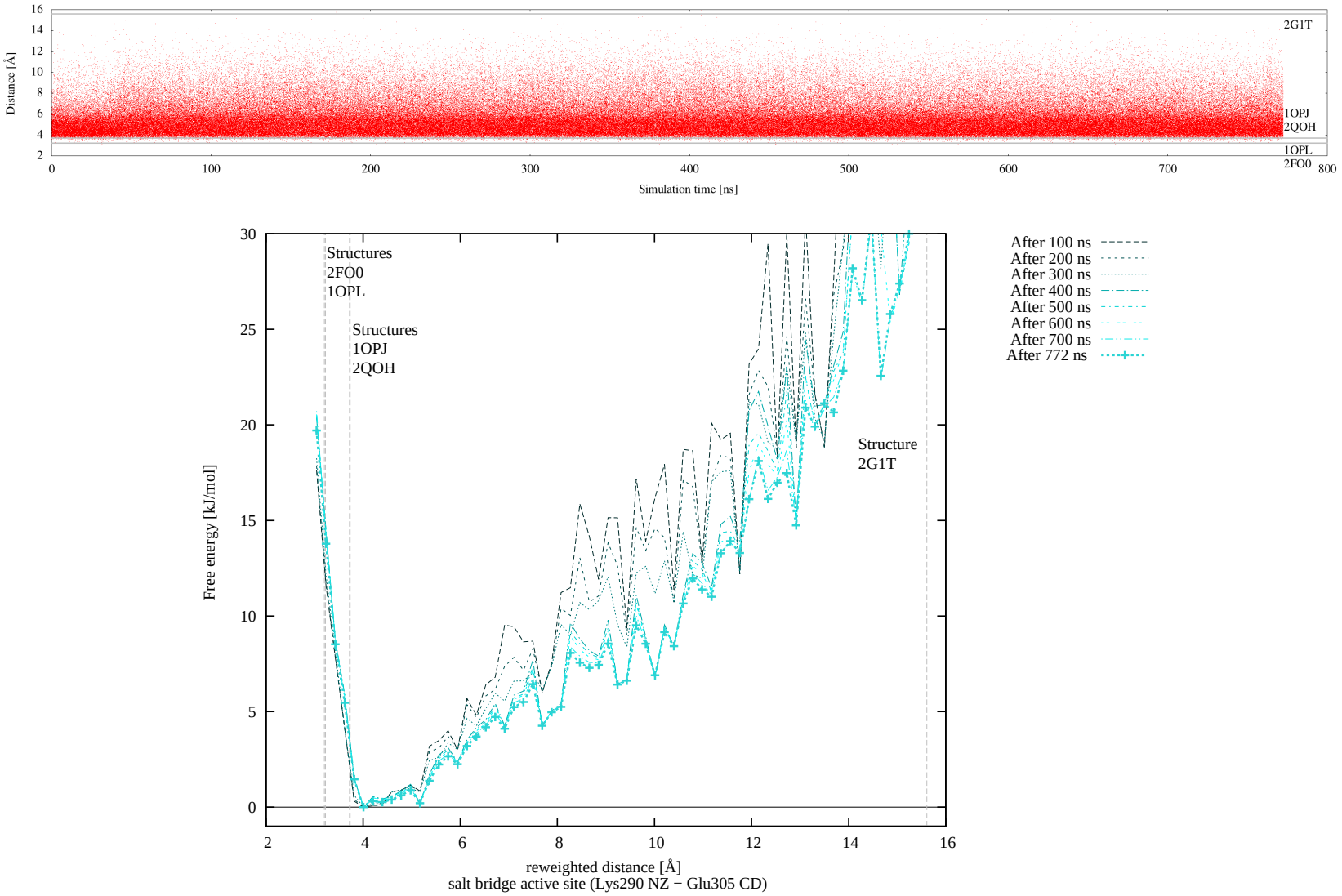

Figure 43: SH2 + Kinase domain. Evolution along simulation time of the extracted distance Lys 290 NZ - Glu 305 CD (active site, active → Abl-like inactive; upper panel) and of its free energy landscape (lower panel) at 310 K.

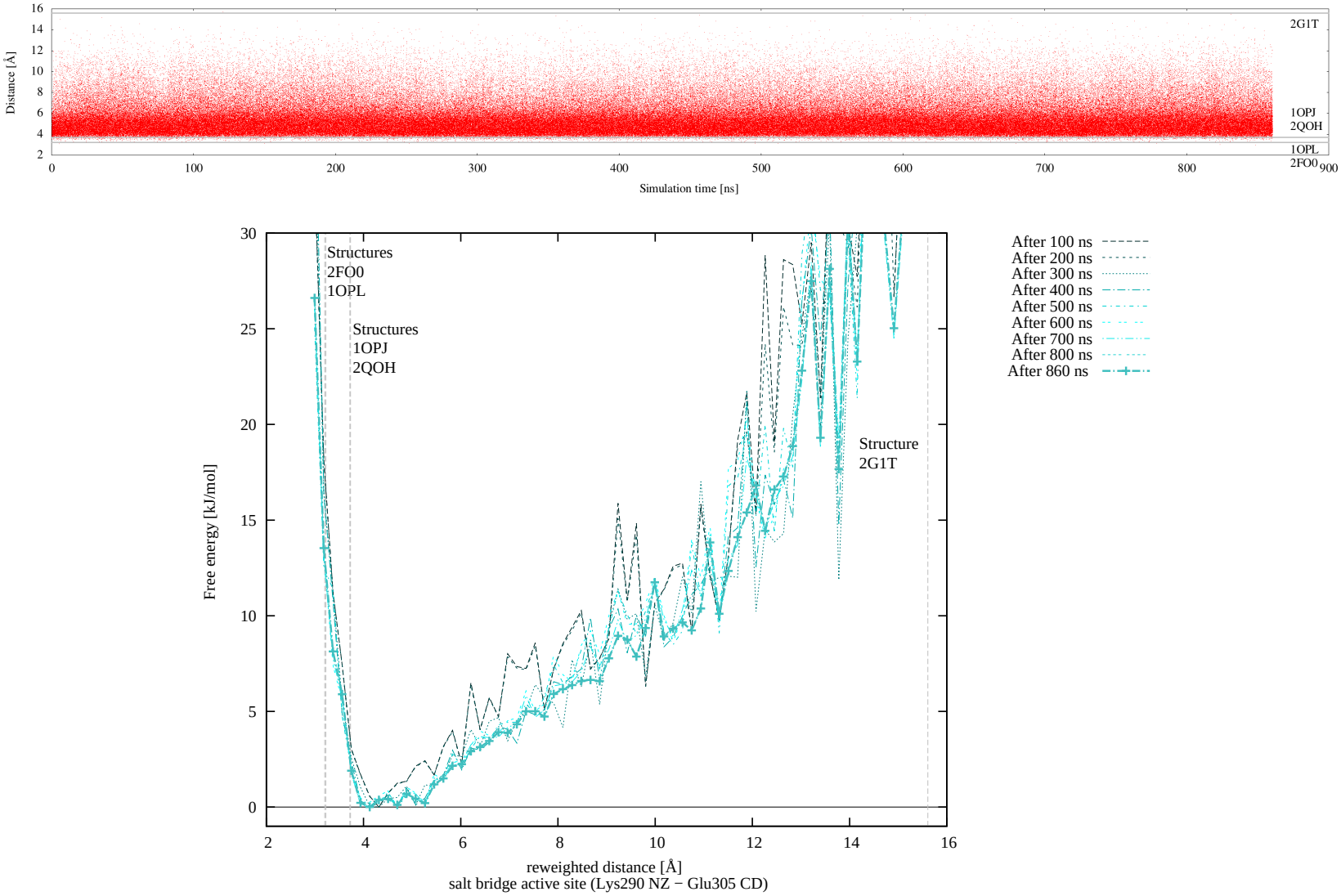

Figure 44: SH3 + SH2 + Kinase domain. Evolution along simulation time of the extracted distance Lys 290 NZ - Glu 305 CD (active site, active → Abl-like inactive; upper panel) and of its free energy landscape (lower panel) at 310 K.

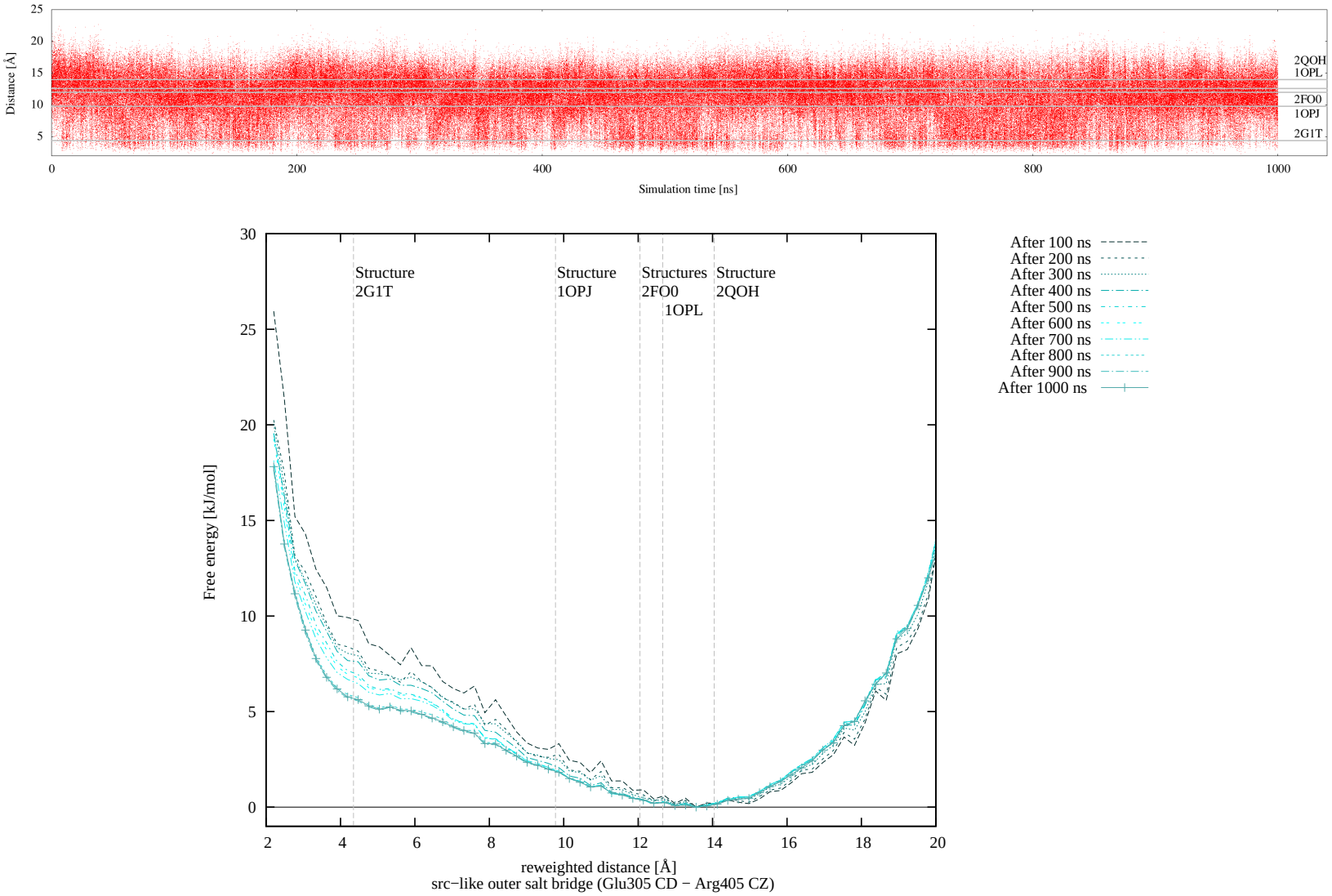

Figure 45: Kinase domain. Evolution along simulation time of the extracted distance Glu 305 CD - Arg 405 CZ (external salt bridge, Src-like inactive $\rightarrow$ any other state; upper panel) and of its free energy landscape (lower panel) at 310 K.

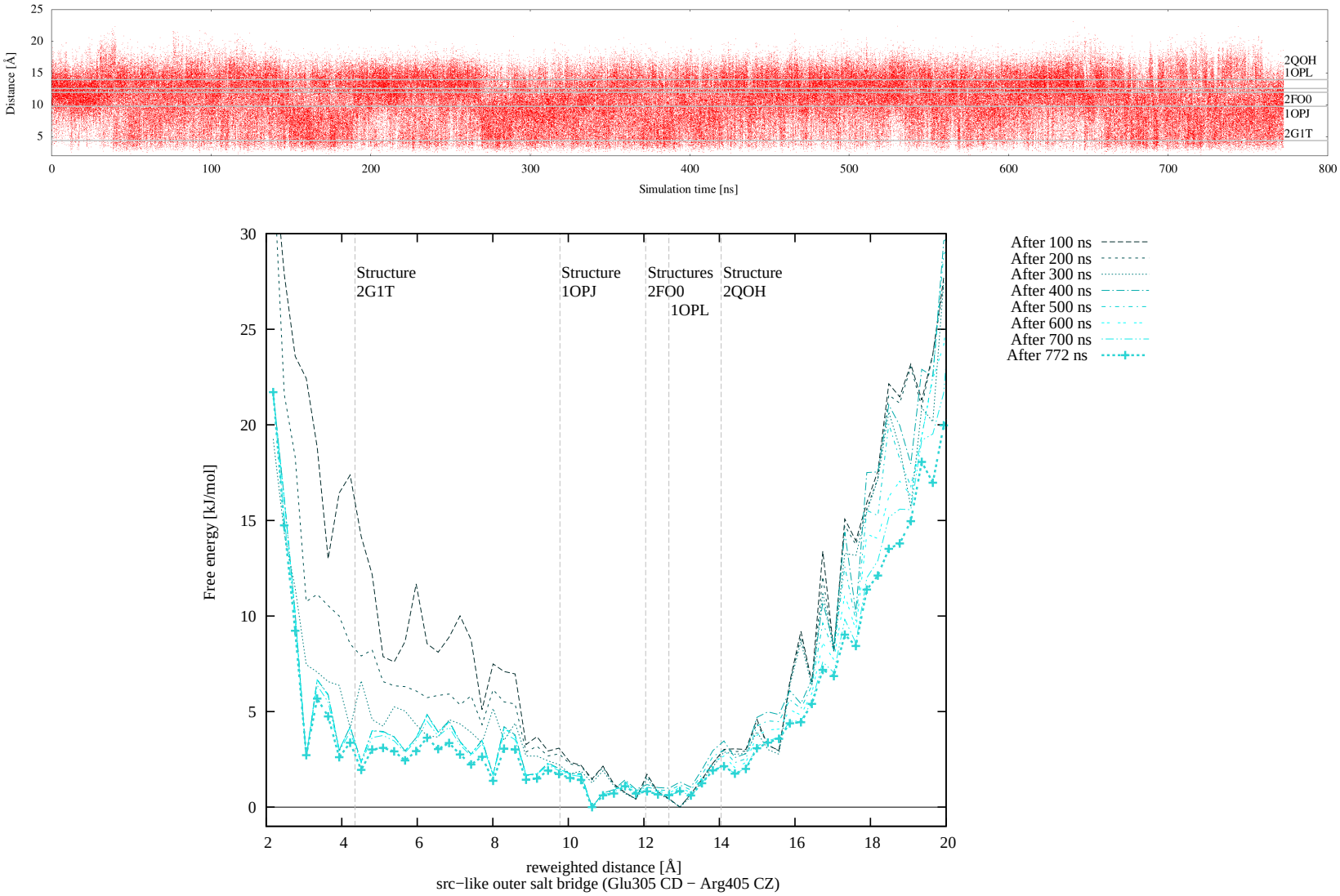

Figure 46: SH2 + Kinase domain. Evolution along simulation time of the extracted distance Glu 305 CD - Arg 405 CZ (external salt bridge, Src-like inactive $\rightarrow$ any other state; upper panel) and of its free energy landscape (lower panel) at 310 K.

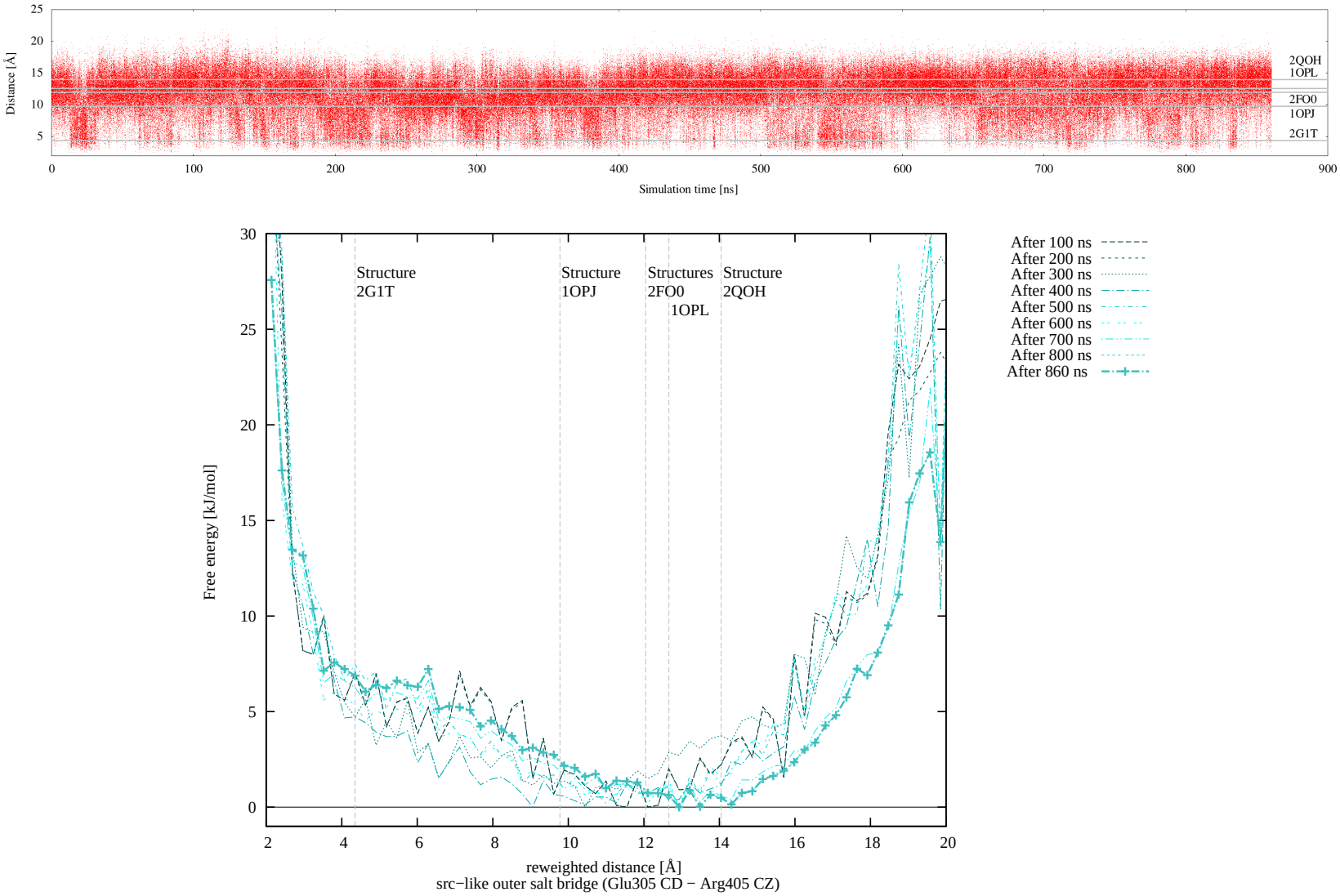

Figure 47: SH3 + SH2 + Kinase domain. Evolution along simulation time of the extracted distance Glu 305 CD - Arg 405 CZ (external salt bridge, Src-like inactive $\rightarrow$ any other state; upper panel) and of its free energy landscape (lower panel) at 310 K.

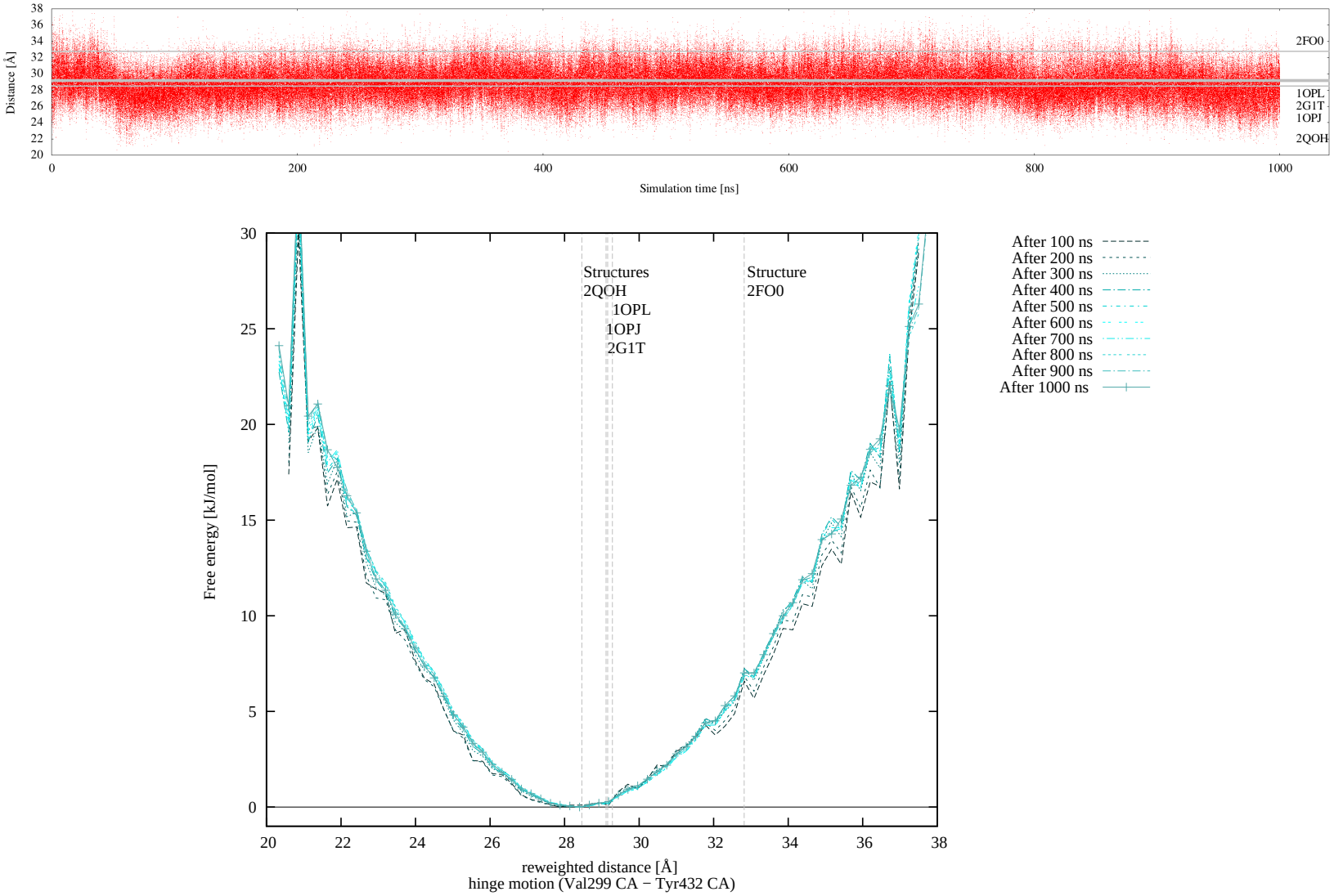

Figure 48: Kinase domain. Evolution along simulation time of the extracted hinge motion measure (distance Val 299 CA - Tyr 432 CA, upper panel) and of its free energy landscape (lower panel) at 310 K.

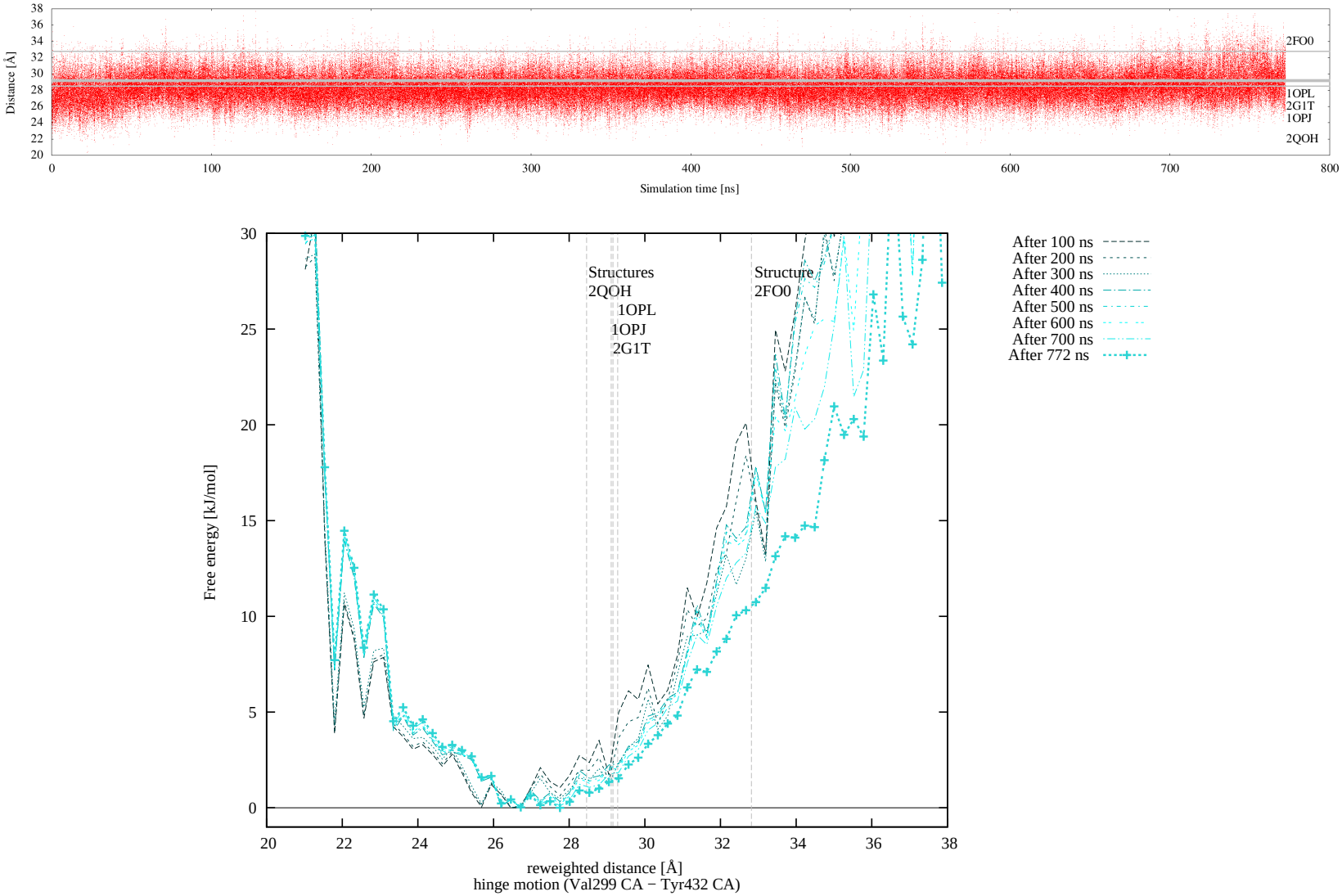

Figure 49: SH2 + Kinase domain. Evolution along simulation time of the extracted hinge motion measure (distance Val 299 CA - Tyr 432 CA, upper panel) and of its free energy landscape (lower panel) at 310 K.

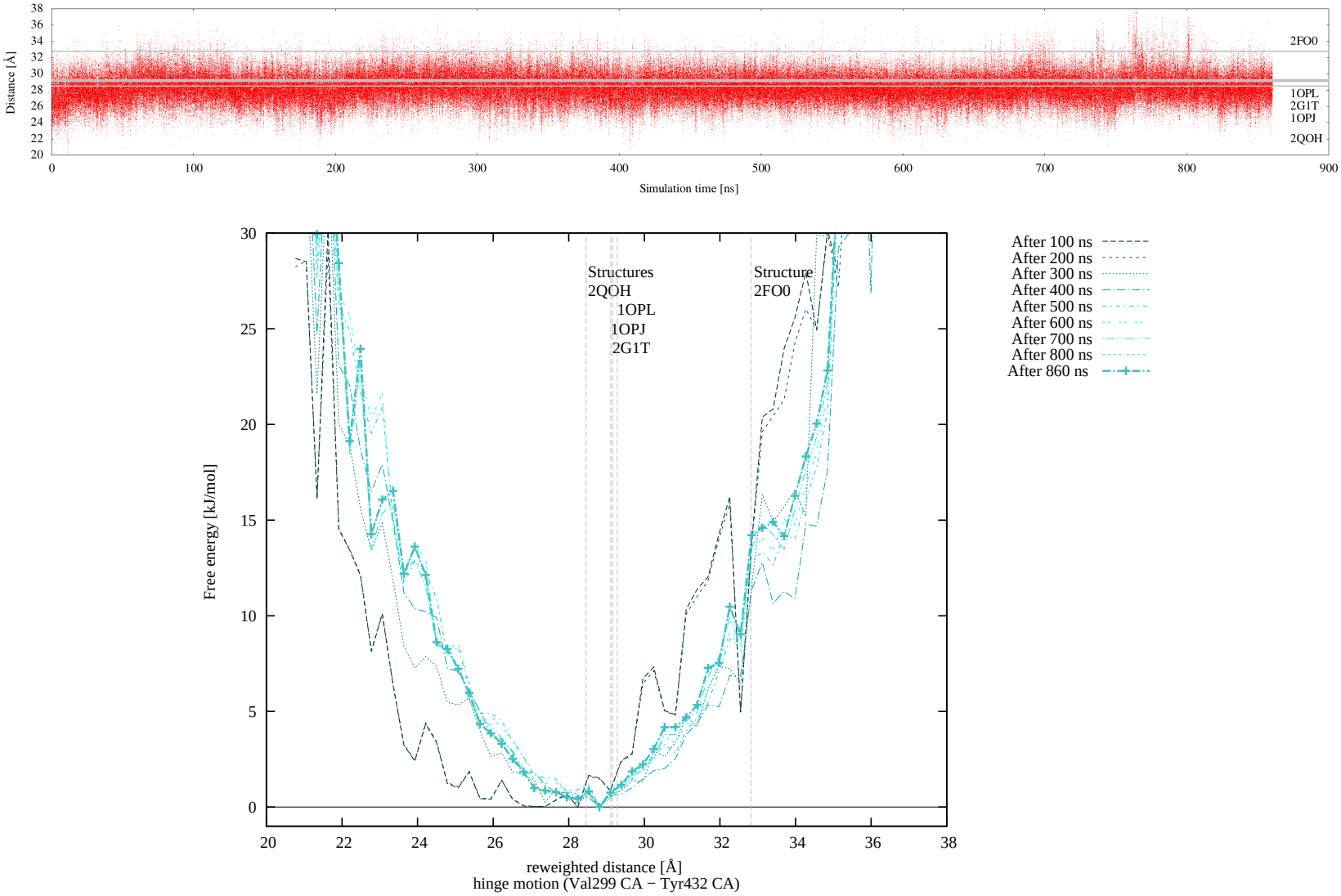

Figure 50: SH3 + SH2 + Kinase domain. Evolution along simulation time of the extracted hinge motion measure (distance Val 299 CA - Tyr 432 CA, upper panel) and of its free energy landscape (lower panel) at 310 K.

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

[1] J. R. Knowles. Enzyme catalysis: not different, just better. *Nature*, 350(6314):121–124, Mar 1991. (Cited on page 1.)

[2] J. A. McCammon, B. R. Gelin, M. Karplus, and P. G. Wolynes. The hinge-bending mode in lysozyme. *Nature*, 262(5566):325–326, Jul 1976. (Cited on page 1.)

[3] J. A. McCammon, B. R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, Jun 1977. (Cited on page 1.)

[4] D. Frenkel and B. Smit. *Understanding Molecular Simulation*. Academic Press, Inc., Orlando, FL, USA, 2nd edition, 2001. (Cited on page 2.)

[5] R. A. V. Etten. Cycling, stressed-out and nervous: cellular functions of c-abl. *Trends in Cell Biology*, 9(5):179 – 186, 1999. (Cited on pages 3, 33, and 42.)

[6] J. J. Wu, H. Phan, and K. S. Lam. Comparison of the intrinsic kinase activity and substrate specificity of c-abl and bcr-abl. *Bioorganic & Medicinal Chemistry Letters*, 8(17):2279 – 2284, 1998. (Cited on page 3.)

[7] P. C. Nowell and D. A. Hungerford. A minute chromosome in human chronic granulocytic leukemia. *Science*, 132(3438):1457–1508, Nov 1960. (Cited on page 3.)

[8] C.-H. Pui, M. V. Relling, and J. R. Downing. Acute lymphoblastic leukemia. *New England Journal of Medicine*, 350(15):1535–1548, 2004. PMID: 15071128. (Cited on page 3.)

[9] O. Hantschel and G. Superti-Furga. Regulation of the c-abl and bcr-abl tyrosine kinases. *Nat Rev Mol Cell Biol*, 5(1):33–44, Jan 2004. (Cited on page 3.)

[10] A. M. Pendergast, M. L. Gishizky, M. H. Havlik, and O. N. Witte. Sh1 domain autophosphorylation of p210 bcr/abl is required for transforma-

tion but not growth factor independence. *Molecular and Cellular Biology*, 13(3):1728–1736, 1993. (Cited on page 3.)

[11] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002. (Cited on pages 3, 4, and 34.)

[12] M. Huse and J. Kuriyan. The conformational plasticity of protein kinases. *Cell*, 109(3):275 – 282, 2002. (Cited on pages 4 and 37.)

[13] J. C. Venter, M. D. Adams, E. W. Myers, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001. (Cited on page 4.)

[14] B. J. Druker and N. B. Lydon. Lessons learned from the development of an abl tyrosine kinase inhibitor for chronic myelogenous leukemia. *The Journal of Clinical Investigation*, 105(1):3–7, 1 2000. (Cited on page 4.)

[15] B. J. Druker, S. Tamura, E. Buchdunger, et al. Effects of a selective inhibitor of the abl tyrosine kinase on the growth of bcr-abl positive cells. *Nat Med*, 2(5):561–566, May 1996. (Cited on page 4.)

[16] B. J. Druker, M. Talpaz, D. J. Resta, et al. Efficacy and safety of a specific inhibitor of the bcr-abl tyrosine kinase in chronic myeloid leukemia. *New England Journal of Medicine*, 344(14):1031–1037, 2001. PMID: 11287972. (Cited on page 4.)

[17] G. Saglio, D.-W. Kim, S. Issaragrisil, et al. Nilotinib versus imatinib for newly diagnosed chronic myeloid leukemia. *New England Journal of Medicine*, 362(24):2251–2259, 2010. PMID: 20525993. (Cited on page 4.)

[18] H. Kantarjian, N. P. Shah, A. Hochhaus, et al. Dasatinib versus imatinib in newly diagnosed chronic-phase chronic myeloid leukemia. *New England Journal of Medicine*, 362(24):2260–2270, 2010. PMID: 20525995. (Cited on page 4.)

[19] F. Grebien, O. Hantschel, J. Wojcik, et al. Targeting the sh2-kinase interface in bcr-abl inhibits leukemogenesis. *Cell*, 147(2):306–319, 10 2011. (Cited on page 5.)

[20] M. E. M. Noble, J. A. Endicott, and L. N. Johnson. Protein kinase inhibitors: Insights into drug design from structure. *Science*, 303(5665):1800–1805, 2004. (Cited on page 5.)

[21] J. Durrant and J. A. McCammon. Molecular dynamics simulations and drug discovery. *BMC Biology*, 9(1):71, 2011. (Cited on page 6.)

[22] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, et al. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010. (Cited on page 6.)

[23] Y. Shan, E. T. Kim, M. P. Eastwood, et al. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24):9181–9183, 2011. (Cited on page 6.)

[24] Y. Shan, M. Eastwood, X. Zhang, et al. Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. *Cell*, 149(4):860 – 870, 2012. (Cited on page 6.)

[25] A. Laio and F. L. Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, 71(12):126601, 2008. (Cited on pages 6 and 11.)

[26] M. Bonomi, D. Branduardi, G. Bussi, et al. Plumed: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications*, 180(10):1961 – 1972, 2009. (Cited on pages 6 and 13.)

[27] Y. Duan, C. Wu, S. Chowdhury, et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry*, 24(16):1999–2012, 2003. (Cited on page 9.)

[28] V. Hornak, R. Abel, A. Okur, et al. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725, 2006. (Cited on page 9.)

[29] R. B. Best and G. Hummer. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *The Journal of Physical Chemistry B*, 113(26):9004–9015, 2009. PMID: 19514729. (Cited on pages 9 and 21.)

[30] K. Lindorff-Larsen, S. Piana, K. Palmo, et al. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958, 2010. (Cited on pages 9 and 21.)

[31] A. D. MacKerell, D. Bashford, Bellott, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998. (Cited on page 9.)

[32] A. D. Mackerell, M. Feig, and C. L. Brooks. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry*, 25(11):1400–1415, 2004. (Cited on page 9.)

[33] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. *Biophysical journal*, volume 100, chapter How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization?, pages L47–L49. Cell Press, May 2011. (Cited on pages 10 and 53.)

[34] D. A. Case, T. E. Cheatham, T. Darden, et al. The amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, 2005. (Cited on page 10.)

[35] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, et al. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983. (Cited on page 10.)

[36] L. Kalé, R. Skeel, M. Bhandarkar, et al. Namd2: Greater scalability for parallel molecular dynamics. *Journal of Computational Physics*, 151(1):283 – 312, 1999. (Cited on page 10.)

[37] J. C. Phillips, R. Braun, W. Wang, et al. Scalable molecular dynamics with namd. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005. (Cited on page 10.)

[38] H. Berendsen, D. van der Spoel, and R. van Drunen. Gromacs: A message-passing parallel molecular dynamics implementation. *Computer*

Physics Communications, 91(1âĂŞ3):43 – 56, 1995. (Cited on pages 10 and 18.)

[39] E. Lindahl, B. Hess, and D. van der Spoel. Gromacs 3.0: a package for molecular simulation and trajectory analysis. *Molecular modeling annual*, 7(8):306–317, 2001. (Cited on pages 10 and 18.)

[40] D. Van Der Spoel, E. Lindahl, B. Hess, et al. Gromacs: Fast, flexible, and free. *Journal of Computational Chemistry*, 26(16):1701–1718, 2005. (Cited on pages 10 and 18.)

[41] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008. (Cited on pages 10 and 18.)

[42] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1-2):141 – 151, 1999. (Cited on page 11.)

[43] M. M. Seibert, A. Patriksson, B. Hess, and D. van der Spoel. Reproducible polypeptide folding and structure prediction using molecular dynamics simulations. *Journal of Molecular Biology*, 354(1):173 – 183, 2005. (Cited on page 11.)

[44] A. Laio and M. Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002. (Cited on page 11.)

[45] C. Theodoropoulos, Y.-H. Qian, and I. G. Kevrekidis. "Coarse" stability and bifurcation analysis using time-steppers: A reaction-diffusion example. *Proceedings of the National Academy of Sciences*, 97(18):9840–9843, 2000. (Cited on page 11.)

[46] I. G. Kevrekidis, C. W. Gear, and G. Hummer. Equation-free: The computer-aided analysis of complex multiscale systems. *AIChE Journal*, 50(7):1346–1355, 2004. (Cited on page 11.)

[47] T. Huber, A. Torda, and W. Gunsteren. Local elevation: A method for improving the searching properties of molecular dynamics simulation.

Journal of Computer-Aided Molecular Design, 8(6):695–708, 1994. (Cited on page 11.)

[48] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation - Umbrella sampling. *Journal of Computational Physics*, 23:187–199, February 1977. (Cited on page 11.)

[49] J. F. Dama, M. Parrinello, and G. A. Voth. Well-tempered metadynamics converges asymptotically. *Phys. Rev. Lett.*, 112:240602, Jun 2014. (Cited on page 12.)

[50] A. Barducci, G. Bussi, and M. Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.*, 100:020603, Jan 2008. (Cited on page 13.)

[51] G. Bussi, A. Laio, and M. Parrinello. Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.*, 96:090601, Mar 2006. (Cited on page 13.)

[52] G. Bussi, F. L. Gervasio, A. Laio, and M. Parrinello. Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics. *Journal of the American Chemical Society*, 128(41):13435–13441, 2006. PMID: 17031956. (Cited on pages 13, 41, and 55.)

[53] D. Branduardi, F. L. Gervasio, and M. Parrinello. From A to B in free energy space. *The Journal of Chemical Physics*, 126(5):–, 2007. (Cited on page 14.)

[54] M. Bonomi, A. Barducci, and M. Parrinello. Reconstructing the equilibrium boltzmann distribution from well-tempered metadynamics. *Journal of Computational Chemistry*, 30(11):1615–1621, 2009. (Cited on pages 15 and 56.)

[55] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21(3):167–195, 1995. (Cited on page 16.)

[56] R. B. Laughlin, D. Pines, J. Schmalian, B. P. Stojković, and P. Wolynes. The middle way. *Proceedings of the National Academy of Sciences*, 97(1):32–37, 2000. (Cited on page 16.)

[57] H. Taketomi, Y. Ueda, and N. Gō. Studies on protein folding, unfolding and fluctuations by computer simulation. *International Journal of Peptide and Protein Research*, 7(6):445–459, 1975. (Cited on pages 16 and 23.)

[58] N. Gō. Theoretical studies of protein folding. *Annual Review of Biophysics and Bioengineering*, 12(1):183–210, 1983. (Cited on page 16.)

[59] C. Clementi. Coarse-grained models of protein folding: toy models or predictive tools? *Current Opinion in Structural Biology*, 18(1):10 – 15, 2008. (Cited on page 16.)

[60] V. Tozzini. Coarse-grained models for proteins. *Current Opinion in Structural Biology*, 15(2):144 – 150, 2005. (Cited on page 16.)

[61] C. Clementi, H. Nymeyer, and J. N. Onuchic. Topological and energetic factors: what determines the structural details of the transition state ensemble and en-route intermediates for protein folding? an investigation for small globular proteins. *Journal of Molecular Biology*, 298(5):937 – 953, 2000. (Cited on page 16.)

[62] R. D. Hills and C. L. Brooks. Insights from coarse-grained Gō models for protein folding and dynamics. *International Journal of Molecular Sciences*, 10(3):889–905, 2009. (Cited on page 17.)

[63] P. C. Whitford, J. K. Noel, S. Gosavi, et al. An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields. *Proteins: Structure, Function, and Bioinformatics*, 75(2):430 – 441, 2009. (Cited on pages 17 and 18.)

[64] L. Wu, J. Zhang, M. Qin, F. Liu, and W. Wang. Folding of proteins with an all-atom g-model. *The Journal of Chemical Physics*, 128(23):–, 2008. (Cited on pages 17 and 19.)

[65] A. Schug, M. Weigt, J. N. Onuchic, T. Hwa, and H. Szurmant. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proceedings of the National Academy of Sciences*, 106(52):22124–22129, 2009. (Cited on page 17.)

[66] T. V. Pogorelov and Z. Luthey-Schulten. Variations in the fast folding rates of the λ-repressor: A hybrid molecular dynamics study. *Biophysical journal*, 87(1):207–214, Jul 2004. (Cited on page 17.)

[67] J. H. Meinke and U. H. E. Hansmann. Protein simulations combining an all-atom force field with a go term. *Journal of Physics: Condensed Matter*, 19(28):285215, 2007. (Cited on page 17.)

[68] L. Sutto, I. Mereu, and F. L. Gervasio. A hybrid all-atom structure-based model for protein folding and large scale conformational transitions. *Journal of Chemical Theory and Computation*, 7(12):4208–4217, 2011. (Cited on pages 17 and 30.)

[69] J. K. Noel, P. C. W, K. Y. Sanbonmatsu, and J. N. Onuchic. Smogctbp: simplified deployment of structure-based models in gromacs. *Nucleic Acids Research*, 2010. (Cited on page 18.)

[70] V. P. Grantcharova and D. Baker. Folding dynamics of the src sh3 domain. *Biochemistry*, 36(50):15685–15692, 1997. (Cited on pages 23, 27, and 30.)

[71] I. V. Kalgin, M. Karplus, and S. F. Chekmarev. Folding of a sh3 domain: Standard and "hydrodynamic"analyses. *The Journal of Physical Chemistry B*, 113(38):12759–12772, 2009. (Cited on page 23.)

[72] W. Xu, S. C. Harrison, and M. J. Eck. Three-dimensional structure of the tyrosine kinase c-src. *Nature*, 385:595 – 602, 1997. (Cited on page 23.)

[73] H. Kaya and H. S. Chan. Polymer principles of protein calorimetric two-state cooperativity. *Proteins: Structure, Function, and Bioinformatics*, 40(4):637–661, 2000. (Cited on page 25.)

[74] A. M. Ferrenberg and R. H. Swendsen. Optimized monte carlo data analysis. *Phys. Rev. Lett.*, 63:1195–1198, Sep 1989. (Cited on page 25.)

[75] V. P. Grantcharova, D. S. Riddle, J. V. Santiago, and D. Baker. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src sh3 domain. *Nat Struct Mol Biol*, 5(8):714–720, Aug 1998. (Cited on pages 29 and 30.)

[76] J. C. McKnight, D. S. Doering, P. T. Matsudaira, and P. S. Kim. A thermostable 35-residue subdomain within villin headpiece. *Journal of Molecular Biology*, 260(2):126 – 134, 1996. (Cited on page 30.)

[77] Y. Duan and P. A. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282(5389):740–744, 1998. (Cited on page 30.)

[78] L. Vugmeyster, D. Ostrovsky, A. Khadjinova, et al. Slow motions in the hydrophobic core of chicken villin headpiece subdomain and their contributions to configurational entropy and heat capacity from solid-state deuteron nmr measurements. *Biochemistry*, 50(49):10637–10646, 2011. (Cited on page 30.)

[79] A. L. Serrano, O. Bilsel, and F. Gai. Native state conformational heterogeneity of hp35 revealed by time-resolved fret. *The Journal of Physical Chemistry B*, 116(35):10631–10638, 2012. (Cited on page 30.)

[80] E. Shtivelman, B. Lifshitz, R. P. Gale, B. A. Roe, and E. Canaani. Alternative splicing of RNAs transcribed from the human abl gene and from the bcr-abl fused gene. *Cell*, 47(2):277 – 284, 1986. (Cited on page 33.)

[81] N. M. Levinson, O. Kuchment, K. Shen, et al. A src-like inactive conformation in the abl tyrosine kinase domain. *PLoS Biol*, 4(5):e144, 05 2006. (Cited on pages 39, 40, 41, and 44.)

[82] M. A. Seeliger, B. Nagar, F. Frank, et al. c-src binds to the cancer drug imatinib with an inactive abl/c-kit conformation and a distributed thermodynamic penalty. *Structure*, 15(3):299 – 311, 2007. (Cited on page 39.)

[83] S. Lovera, L. Sutto, R. Boubeva, et al. The different flexibility of c-src and c-abl kinases regulates the accessibility of a druggable inactive conformation. *Journal of the American Chemical Society*, 134(5):2496–2499, 2012. (Cited on pages 40 and 41.)

[84] Y. Shan, M. A. Seeliger, M. P. Eastwood, et al. A conserved protonation-dependent switch controls drug binding in the abl kinase. *Proceedings of the National Academy of Sciences*, 106(1):139–144, 2009. (Cited on page 40.)

[85] Y.-L. Lin, Y. Meng, W. Jiang, and B. Roux. Explaining why gleevec is a specific and potent inhibitor of abl kinase. *Proceedings of the National Academy of Sciences*, 2013. (Cited on pages 40 and 41.)

[86] A. Aleksandrov and T. Simonson. Molecular dynamics simulations show that conformational selection governs the binding preferences of imatinib for several tyrosine kinases. *Journal of Biological Chemistry*, 285(18):13807–13815, 2010. (Cited on page 41.)

[87] H. Pluk, K. Dorey, and G. Superti-Furga. Autoinhibition of c-abl. *Cell*, 108(2):247 – 259, 2002. (Cited on page 42.)

[88] B. Nagar, O. Hantschel, M. A. Young, et al. Structural basis for the autoinhibition of c-abl tyrosine kinase. *Cell*, 112(6):859 – 871, 2003. (Cited on pages 42, 43, 44, and 45.)

[89] I. Sadowski, J. C. Stone, and T. Pawson. A noncatalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of fujinami sarcoma virus p130gag-fps. *Molecular and Cellular Biology*, 6(12):4396–4408, 1986. (Cited on page 42.)

[90] S. Zhou, S. E. Shoelson, M. Chaudhuri, et al. {SH2} domains recognize specific phosphopeptide sequences. *Cell*, 72(5):767 – 778, 1993. (Cited on page 42.)

[91] P. Filippakopoulos, S. Müller, and S. Knapp. SH2domains: modulators of nonreceptor tyrosine kinase activity. *Current Opinion in Structural Biology*, 19(6):643 – 649, 2009. (Cited on page 42.)

[92] G. Waksman, D. Kominos, S. C. Robertson, et al. Crystal structure of the phosphotyrosine recognition domain sh2 of v-src complexed with tyrosine-phosphorylated peptides. *Nature*, 358:646 – 653, 1992. (Cited on page 42.)

[93] P. Filippakopoulos, M. Kofler, O. Hantschel, et al. Structural coupling of sh2-kinase domains links fes and abl substrate recognition and kinase activation. *Cell*, 134(5):793 – 803, 2008. (Cited on pages 42 and 46.)

[94] B. Nagar, W. G. Bornmann, P. Pellicena, et al. Crystal structures of the kinase domain of c-abl in complex with the small molecule inhibitors pd173955 and imatinib (sti-571). *Cancer Research*, 62(9):4236 – 4243, 2002. (Cited on page 42.)

[95] H.-J. Nam, W. G. Haser, T. M. Roberts, and C. A. Frederick. Intramolecular interactions of the regulatory domains of the bcr-abl kinase reveal a novel control mechanism. *Structure*, 4(9):1105 – 1114, 1996. (Cited on page 42.)

[96] T. Pawson and J. D. Scott. Signaling through scaffold, anchoring, and adaptor proteins. *Science*, 278(5346):2075–2080, 1997. (Cited on page 43.)

[97] B. Nagar, O. Hantschel, M. Seeliger, et al. Organization of the sh3-sh2 unit in active and inactive forms of the c-abl tyrosine kinase. *Molecular Cell*, 21(6):787 – 798, 2006. (Cited on pages 43, 45, and 46.)

[98] S. Chen, S. Brier, T. E. Smithgall, and J. R. Engen. The abl sh2-kinase linker naturally adopts a conformation competent for sh3 domain binding. *Protein Science*, 16(4):572–581, 2007. (Cited on page 43.)

[99] W. Xu, A. Doshi, M. Lei, M. J. Eck, and S. C. Harrison. Crystal structures of c-src reveal features of its autoinhibitory mechanism. *Molecular cell*, 3(5):629–638, 1999. (Cited on page 44.)

[100] T. Schindler, F. Sicheri, A. Pico, et al. Crystal structure of hck in complex with a src family-selective tyrosine kinase inhibitor. *Molecular cell*, 3(5):639–648, 1999. (Cited on page 44.)

[101] R. E. Iacob, T. Pene-Dumitrescu, J. Zhang, et al. Conformational disturbance in abl kinase upon mutation and deregulation. *Proceedings of the National Academy of Sciences*, 106(5):1386–1391, 2009. (Cited on page 44.)

[102] T. Zhou, L. Parillon, F. Li, et al. Crystal structure of the t315i mutant of abl kinase. *Chemical Biology & Drug Design*, 70(3):171–181, 2007. (Cited on page 44.)

[103] O. Hantschel, B. Nagar, S. Guettler, et al. A myristoyl/phosphotyrosine switch regulates c-abl. *Cell*, 112(6):845 – 857, 2003. (Cited on page 46.)

[104] V. B. Chen, W. B. Arendall, III, J. J. Headd, et al. *MolProbity*: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D*, 66(1):12–21, Jan 2010. (Cited on page 53.)

[105] A. D. MacKerell, M. Feig, and C. L. Brooks. Improved treatment of the protein backbone in empirical force fields. *Journal of the American Chemical Society*, 126(3):698–699, 2004. (Cited on page 53.)

[106] R. B. Best and G. Hummer. Optimized molecular dynamics force fields applied to the helixâĹŠcoil transition of polypeptides. *The Journal of Physical Chemistry B*, 113(26):9004–9015, 2009. PMID: 19514729. (Cited on page 53.)

[107] F. Avbelj, S. G. Grdadolnik, J. Grdadolnik, and R. L. Baldwin. Intrinsic backbone preferences are fully present in blocked amino acids. *Proceedings of the National Academy of Sciences of the United States of America*, 103(5):1272–1277, 2006. (Cited on page 53.)

[108] N. Homeyer, A. Horn, H. Lanig, and H. Sticht. Amber force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *Journal of Molecular Modeling*, 12(3):281–289, 2006. (Cited on page 53.)

[109] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996. (Cited on page 53.)

[110] The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC. 2010. (Cited on page 53.)

[111] N. Guex, M. C. Peitsch, and T. Schwede. Automated comparative protein structure modeling with swiss-model and swiss-pdbviewer: A historical perspective. *ELECTROPHORESIS*, 30(S1):S162–S173, 2009. (Cited on page 53.)

[112] G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95 – 99, 1963. (Cited on page 55.)

CURRICULUM VITAE

Ilaria Mereu

EDUCATION

Universidad Autónoma de Madrid (UAM), Madrid (Spain)
Doctor of Philosophy Candidate in Biophysics present

University of Cagliari, Cagliari (Italy)
Master of Science in Theoretical Physics of Condensed Matter July 2009
Magna cum Laude

University of Cagliari, Cagliari (Italy)
Bachelor of Science in Physics October 2005
Magna cum Laude

RESEARCH EXPERIENCE

La Caixa-CNIO Predoctoral Fellow / UAM graduate student 2009 - present
Former Computational Biophysics Junior Group, Spanish National Cancer Research Center (CNIO), Madrid
Dissertation:"Development of an hybrid computational method for the study of large-scale conformational transitions in proteins"
Advisors: Professor Francesco L. Gervasio, Ph.D. and Professor Alfonso Valencia, Ph.D.

Intern - Master Thesis project development December 2008 - June 2009
Sardinian Laboratory for Computational Materials Science (SLACS), Cagliari, Italy
Thesis title: "Improving an all-atom parametrization protocol for medicinal chemistry".
Advisors: A. V. Vargiu, Ph.D., Prof. P. Ruggerone, Ph.D.

Intern - Erasmus program at J. Fourier University, Grenoble April 2006 - July 2006
Institute of Structural Biology (IBS), Grenoble, France
Advisor: Aline Thomas, Ph.D.

Intern - Bachelor thesis project development July 2005 - October 2005
Italian National Institute of Nuclear Physics (INFN), Cagliari, Italy
Thesis title: "Primordial transverse impulse effects in the production of D^0 in ν-nucleon charged current

collisions".
Advisor: Professor B. Saitta, Ph.D.

PUBLICATIONS AND POSTERS

Sutto L, **Mereu I**, Gervasio FL (2011).
A Hybrid All-Atom Structure-Based Model for Protein Folding and Large Scale Conformational Transitions. J Chem Theory Comput 7, 4208-4217.

· 27/2/2012, Biophysical Society 56th Annual Meeting, San Diego, USA
Mereu I, Sutto L, Gervasio FL. *Dynamics of Large-Scale Protein Conformational Transition and Docking Events using a Hybrid All-atom Structure Based Model.*
· 17-19/9/2012, CNIO Frontiers Meeting "Allosteric Regulation of Cell Signaling", Madrid, Spain
Mereu I, Sutto L, Gervasio FL. *Free energy of the interdomain assembly of SH2 and catalytic domain in c-Abl combining metadynamics and all-atom structure-based modeling.*
· 17/2/2014, Biophysical Society 58th Annual Meeting, San Francisco, USA
Mereu I, Sutto L, Gervasio FL. *The free energy contributions of SH3 and SH2 in c-Abl 1b autoinhibition mechanism via a computational structure-based model.*

Language skills: Italian (native), English (fluent), Spanish (fluent), French (intermediate).